CRISPR IMMUNE DIVERSITY IN SIMULATED AND NATURAL MICROBIAL POPULATIONS

BY

WHITNEY E. ENGLAND

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Microbiology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Associate Professor Rachel J. Whitaker, Chair
Professor Isaac K. O. Cann
Professor Gary J. Olsen
Associate Professor Joanna L. Shisler

ABSTRACT

Coevolution between microbes and their viruses influences the trajectories of these communities through gene transfer and predation. When these communities are a part of the human microbiome, these interactions can also have significant impacts on the health of the human host. The CRISPR adaptive immune system is one of the ways in which microbes defend against viral infection, and it also holds a record of acquired immunity, allowing us to read a history of microbe-viral interactions. In this work, we examine the emergence, impact, and applications of diverse CRISPR immune alleles in microbial populations. Using a mathematical model of CRISPR-mediated host-virus coevolution to simulate microbial populations, we observe the emergence of multiple coexisting CRISPR alleles in a single population, which we call distributed immunity. We find that distributed immunity is most likely to occur in communities with more potential spacers and relatively low viral mutation rates, and that it is linked to increased stability for the host population, while the viral population is driven to lower densities or even to extinction. To see if this phenomenon is also present in natural microbial populations, we examined CRISPR diversity in two human-associated communities: the vaginal microbiomes of pregnant women and the lung microbiomes of cystic fibrosis patients. To investigate the vaginal microbiome, we developed a network-based methodology to identify and extract CRISPR spacers from all species present in samples taken from pregnant women at high and low risk of preterm birth. This approach yielded over 20 different CRISPR types, with spacer content varying among individuals. Coexisting alleles linked to shifts in the abundance of the matched element were detected in one *Lactobacillus* species in one of the samples, demonstrating the potential of our approach. In the cystic fibrosis lung microbiome, we used this method to identify CRISPRs in four patients infected with the major cystic fibrosis pathogen *Pseudomonas aeruginosa.* Spacer content was completely different between patients, but no variation was detected within a patient. Finally, we examined spacer diversity in a large global dataset of *P. aeruginosa* and used the thousands of spacers identified as a tracking tool to monitor dynamics of viral populations. This approach, which we refer to as prototyping, revealed a panmictic *P. aeruginosa* phage population and holds promise as a tool for tracking mobile elements and personalizing phage therapy treatments.

ACKNOWLEDGMENTS

Writing this dissertation has been an adventure, and it is not the type of adventure one can complete alone. I have been incredibly fortunate to have the personal and professional support of many fine people over the years, and without their invaluable assistance none of you would be reading this document. A full account of everything they have done for me would require an additional volume, but I will do my best to briefly convey their excellence here.

I must start by recognizing my advisor, Dr. Rachel Whitaker. Her scientific passion and relentless drive have been inspirational, and her vision has shaped the way I think about science. Where my project has ended could not be further from where it started, but it has been an exciting journey I am privileged to have taken with you on board. I am also grateful to my committee members, Dr. Isaac Cann, Dr. Gary Olsen, Dr. Joanna Shisler, and Dr. Jim Slauch, for their ideas and guidance, and Dr. Patrick Degnan for his virus-hunting insight.

The Whitaker lab is a special place; together, we have commiserated and celebrated, taken on the rugged landscape of Yellowstone, and helped keep Murphy's in business with "external lab meetings". Thanks to all members past and present for making our lab what it is today. In particular, I am eternally grateful to Dr. Maria Bautista and Dr. Angela Kouris for all the years of friendship, keeping me sane during the writing process, ruthlessly proofreading this manuscript, and providing sustenance just when it was needed most. You are the best friends anyone could ask for. I also thank Dave Krause for his participation in our endlessly entertaining and enlightening conversations.

Every chapter of this thesis has been bolstered by the contributions of my collaborators & co-authors. I thank Dr. Lauren Childs, Dr. Joshua Weitz, and Dr. Mark Young for their roles in developing mathematical models (Chapter 2); Dr. Fang Yang, Dr. Bryan White, Dr. Yan Wei Lim, Dr. Katrine Whiteson, Dr. Forest Rohwer, and Dr. Jeremy Dettman for their willingness to share samples and sequencing data (Chapters 3 & 4); and Ted Kim for his contributions to phage clustering and pan-genome analysis (Chapter 5).

A multitude of staff at the University have also lent their expertise to making this thesis a reality. I thank Dr. Alvaro Hernandez and Chris Wright for all their sequencing expertise, Jeff Haas and all the Life Sciences and IGB IT staff for keeping the servers humming, and administrative staff Diane Tsevelekos, Deb LeBaugh, Shawna Smith, and Connie Scott for keeping it all organized.

This work has been made possible by several generous funders, including the NIH Infection Biology Training Grant, the James R. Beck Graduate Research Fellowship, the Carl R. Woese Institute for Genomic Biology, and the Institute for Advanced Computing Applications and Technologies.

Finally, I owe a debt of gratitude to my family for their constant encouragement throughout this process. I thank my parents for not pushing me in any particular direction, but rather encouraging me to follow the paths I found interesting; it is wandering those paths that has brought me where I am today. I also deeply appreciate the ongoing support from my Champaign family: Dr. Angela Kouris, Dr. Maria Bautista (yes, they deserve to be thanked twice!), Ariana Bravo, Madeline Lopez, Dr. Manuel Ortega, Erin Morris, Dr. Ariane Vartanian, and Buckley "Spacer" Roberts.

Again, thank you all. This adventure would not have been the same without any of you.

TABLE OF CONTENTS

CHAPTER 1

Introduction[1]

Microbe-virus coevolution shapes microbial communities of all types, including environmentally important systems from oceans to acid mine drainage [4,74,75] and medically relevant systems such as the gut microbiome [61,72]. Viruses can act as predators, imposing strong selective pressures on their hosts which shape their evolutionary trajectories. When not lethal, viruses are also a source of new, potentially advantageous genes and a vector for moving genes among organisms. When these microbial hosts are also pathogens to a human host, the importance of these viruses can be amplified as their effects on their microbial hosts in turn impact the trajectory of human disease. Histories of microbe-virus interactions and patterns of immunity become critical for understanding how communities develop.

CRISPRs: A Microbial Adaptive Immune System

There are many antiviral defense mechanisms to be found in the microbial world [53]; one of the most fascinating is the CRISPR (clustered regularly interspaced short palindromic repeats) system. This widespread system is present in the majority of all sequenced bacteria and archaea [49] and functions as an adaptive immune system for microbes which possess it. The CRISPR system is comprised of two main components: arrays of the aforementioned short palindromic repeats interspersed with short DNA fragments known as spacers, and a number of CRISPR-associated (cas) genes which carry out the functions of the system. CRISPR spacers often match the sequence of portions of foreign or mobile genetic elements such as viruses, plasmids, and transposons [63]; the matched portion is known as a protospacer. New spacers can be sampled from genetic elements, possibly using double-strand breaks in DNA that occur during replication [55], and integrated at the leader end of the array [5]. This polar addition of new spacers enables usage of CRISPRs as a historical record of host-virus interactions.

[1] Portions of this chapter were previously published as England, W.E. and Whitaker, R.J. (2013). Evolutionary causes and consequences of diversified CRISPR immune profiles in natural populations. *Biochem. Soc. Trans. 41* (6), 1431–1436, and are reprinted here with permission.

While the genes responsible for spacer acquisition are conserved among the many types of CRISPRs present in microbes, the remaining cas genes vary widely between types, as does the implementation of immunity [58,59]. However; all known CRISPR systems follow the same essential protocol. The repeat-spacer array is transcribed to pre-crRNA and then processed into smaller RNA fragments containing the sequence of a single spacer. These crRNAs are used to target protospacer sequences in DNA, or in some cases RNA [33], and a nuclease is recruited to cleave and inactivate the targeted element [12]. To avoid fatal targeting of the host genome, some types of CRISPRs employ protospacer-adjacent motifs, or PAMs. These short sequences are essential for targeting and are found next to protospacers, but are absent from the spacer-separating repeat sequences, preventing targeting of spacers in the host [64].

In response to the CRISPR threat, targeted elements can evade CRISPR immunity by acquiring random mutations in their protospacers so there is no longer a sufficient match between the spacer and the mutated protospacer [22,79], or through the use of anti-CRISPR proteins present in some viruses, which inhibit CRISPR-cas binding or nuclease recruitment [10,11,67]. The ability for viruses and other elements to evade CRISPR immunity through simple mutation highlights the potential advantage of a host maintaining multiple spacers matching different protospacers in the same virus – a diverse immune repertoire could help prevent viral CRISPR escape.

CRISPR Diversity and Evolution

CRISPR immune profiles in natural microbial populations

Polymorphism and rapid evolution of spacers between direct repeats within the bacterial and archaeal genomes was observed long before the mechanism and function of CRISPRs in adaptive immunity were recognized [48,62]. Polymorphism was assessed by looking at deletions of spacers between repeat sequences whose function for the cell was unknown [41]. For example, fingerprints of these repeated regions were used for typing of mycobacterial strains called spoligotyping (spacer oligonucleotide typing) [29] and to infer the structure of these pathogen populations [50]. Differences in diversity among populations were observed; for example, some mycobacterial outbreaks have a coexisting

diverse set of strains (polyclonal) and others have identical alleles (monoclonal) [50]. The discovery that these repeat arrays within populations were specifically related to bacterial and archaeal immunity [5] transformed the significance of spoligotyping by directly relating diversity of repeat loci into variation in host immunity.

Using PCR to amplify, sequence, and assemble CRISPR repeat-spacer alleles from isolated individuals or directly from environmental samples, several studies have recently investigated the diversity of CRISPRs within microbial populations. These studies have revealed a spectrum of population structures ranging from monoclonal to highly polyclonal. *E. coli* isolates exhibit minimal variation; no new spacers are observed in strains which have diverged in the past 250,000 years, and most observed variation appears to be the result of spacer loss [81,82]. Similarly, *Salmonella* CRISPR loci show variation primarily due to spacer deletion, rather than acquisition of new spacers, and appear to be highly monoclonal [25,77]. Increasing in diversity, nearly clonal populations of *Leptospirillum* in acid mine drainage show extensive diversity at the leader end of the CRISPR array but identity at the trailer end [83]. Populations of *Yersinia pestis,* known for extremely low sequence diversity at other genomic loci [3], show nearly clonal trailer-end spacers with leader-end variation [20,73]. At the other extreme, *Streptococcus thermophilus* exhibits hypervariability in its CRISPR loci; diversity is concentrated at the leader end, but multiple trailer types also exist [43]. Completely distinct CRISPR repeat spacer alleles (at both the leader and the trailer end) have been shown to coexist within a single population of the archaeon *Sulfolobus islandicus* [39]. 39 isolates of *S. islandicus* from a single hot spring sample collected in the year 2000 maintain extensive leader-end diversity, but also contain 8-10 completely different trailer-end alleles at each of three loci at relatively even abundance [39]. A later study involving 120 *S. islandicus* strains taken from the same population ten years later found leader- and trailer-end diversity persisted through time [40]. Why do some microbial populations exhibit extensive CRISPR immune diversity while others do not? What does this difference in diversity between populations tell us about the ongoing coevolutionary dynamic in these populations? What effect does diversity in CRISPR immunity have on evolution of pathogen populations?

Forces that lead to monoclonality in CRISPR immunity

Based on simple Lotka-Volterra [56] dynamics, coevolutionary models predict that at any one time and place a population of hosts would have a single dominant CRISPR allele. Ongoing arms races between viruses and microbial hosts lead to periodic oscillations in immune host genotypes and subsequent selection of viral evasion mutants that can subvert the CRISPR immune surveillance [8]. The prediction is that each oscillation is driven by a selective sweep of an effective immune allele to fixation within a population. If these dynamics are actively ongoing, this model for diversity predicts that different populations of the same archaea or bacteria would have different, monoclonal immune alleles, as they could be at different points in their coevolutionary trajectories. Therefore, if population structure is not well understood, this same evolutionary dynamic could result in apparent diversity of a particular allele when compared among populations that are in fact isolated.

Several other evolutionary, but not coevolutionary, forces could result in monoclonality at the CRISPR locus. Other defense loci such as surface resistance could dominate the host-pathogen dynamic, resulting in clonal CRISPR alleles as the linked resistance locus is swept to fixation within the population. Clonality at a CRISPR locus could also result from demographic history such as bottleneck that would reduce diversity to a single individual allele by chance. Since new CRISPR spacers are added to the leader end, monoclonality at the leader end of the locus has been suggested to result from loss of function of the acquisition machinery in a particular system or the CRISPR system in general within a population.

Forces that lead to polyclonality in CRISPR immunity

Three basic theoretical models of host-pathogen coevolution predict that polymorphism can be generated and maintained within populations [13]. In the first, explicit tradeoffs between resistance or immunity and fitness are required for diversity to emerge [89]. In experimental evolutionary models, such tradeoffs have been observed to promote the coexistence of multiple host or viral genotypes over time [9,14,15,54,86,89]. Associating a cost with CRISPR immunity, these models predict that the maintenance of a CRISPR

system is only adaptive if viral diversity is limited. A second model to promote diversity is negative frequency-dependent selection, where the adaptive benefit of a novel allele decreases as it increases in frequency within a population [19,34,52,65]. In this model, diversity is maintained in microbial populations through negative frequency dependence without explicit tradeoffs between resistance alleles. Finally, spatial structure within populations has been shown to promote diversity. Spatial models have been applied to simulations based on CRISPR immunity to predict that they are essential to maintain diversity [30,31].

In applying these coevolutionary models to explain the CRISPR diversity, it is important to consider two crucial elements of the CRISPR system distinguish it in mechanism and model from previously described systems for host-pathogen coevolution. First, the addition of new spacers to the CRISPR system is "Lamarkian" in that new genotypes are created upon a viral encounter and can be passed on to the next generation [51]. Although the frequency at which new spacers are acquired by hosts has not been well defined, it is believed to be higher than the genome mutation rate, leading to the potential for competing mutations to exist within a population at one time [39]. Second, host cells are not subject to a large fitness drag as consequence of investment in new immune phenotypes. The number of potential immune phenotypes for a population of infected cells is limited only by the number of protospacer sites within each virus, and each is likely to have equal fitness consequences to the host cell.

One additional biological factor that has attracted less attention in terms of its impact on maintaining diversity in natural populations has been the effect of reassortment and recombination of CRISPR loci among individuals within a population. This effect is not dependent upon the CRISPR mechanism of mutation and action, but has been shown to be important for any traits that are under strong selection such as those involve in resistance and immunity. The Red Queen hypothesis [84] states that antagonistically coevolving organisms must continually adapt simply to survive against their ever-changing antagonists. This hypothesis has been applied to explain the maintenance of sexual reproduction; sexual exchange of genetic information increases the creation of novel genotypes and consequently new immunity, resistance and virulence mechanisms

[7,35,46,68]. Similarly, recombination between organisms shuffles existing gene content, allowing for new, possibly advantageous combinations. Recombination is predicted to be especially beneficial when the population is under strong selection, as they are in traits involved in resistance and immunity [6,66]. In a microbial population coevolving with lytic viruses, the viral threat provides strong selection pressure for immunity or resistance in the microbial host. The immunity provided by the spacer of the CRISPR-Cas system makes it likely that horizontal transfer of repeat-spacer arrays and *cas* genes would occur in such populations.

A recent study of a single natural *S. islandicus* population found evidence for rapid recombinatorial reassortment of entire CRISPR loci among strains [40]. Out of a set of 53 natural *S. islandicus* isolates with CRISPR loci where leader-end spacers were observed multiple times, leader-end alleles were found linked to different trailer-end alleles in less than 1% of cases; only four examples of identical spacers shared between loci were identified. By contrast, linkage among the three CRISPR loci present in this population was found to be low, indicating that complete repeat-spacer arrays are reassorted throughout this population [40]. These findings are in line with observations from other studies outside of a single natural population. A comparison of natural *E. coli* isolates found that phylogenetically close strains typically have very similar spacers, but in some cases these strains harbor completely different spacer sets; in over half of cases where the spacers were completely different, the spacers match those of very distant strains, implicating horizontal transfer of spacer arrays [81]. A broad comparison of bacterial 16S rRNA, *cas1*, and direct repeat sequences also found evidence for transfer of the CRISPR locus in its entirety [17]. This shuffling of spacer arrays could benefit the host's battle against its viral antagonists. Reassortment of CRISPR loci can redistribute beneficial antiviral spacers through a population, and it provides another avenue for hosts to acquire different spacers rather than relying solely on leader-end addition. In addition, the reassortment of these loci will also prevent selective sweeps from removing polymorphism within a population. Horizontal transfer of *cas* genes from divergent sources has been observed frequently in microbial populations [27,32,40,42,82]. This could result from a similar mechanism in which horizontal gene flow increases the efficacy of selection on these essential pieces of the CRISPR immunity [57,80]. This

observation suggests that variation in the efficiency of recombination and horizontal gene transfer species might impact the level of polyclonality observed within a microbial population.

Model Systems for Observing CRISPR Diversity

Simulated microbial populations

Integrating mechanistic knowledge of CRISPR immunity, several coevolutionary models have been proposed (reviewed in [38]). Many of these models predict CRISPR immune profiles where there is leader-end diversity but trailer-end clonality. Some models predict that due to the rapid acquisition of new spacers, neutral variation persists at the leader end until selection for a particular spacer causes a selective sweep [37]. An alternative model suggests that since each different spacer confers equal immunity against a given virus, diversity is maintained within a population because each distinct genotype has the same immune phenotype [18]. Although not explicitly investigated as the basis through which diversity is maintained, similar dynamics have been observed in other mathematical models [45]. Together these models suggest that biological parameters such as rates of viral mutation, the number of potential protospacers in the virus genome, and the acquisition of new spacers will result in differences in population structures between different host-pathogen pairs, where some stably maintain a diversity over time [45].

As part of my research, I have demonstrated that this phenomenon emerges within simulated populations and quantified the extent to which this mechanism promotes polymorphism over time. The impact of this type of "distributed immunity" results in stable and increased host populations and unstable viral populations since the advantage of each escape mutation has very little advantage to each viral mutant.

Natural microbial communities of the human microbiome

To fully realize the impact of dynamics observed in simulated populations, these observations must be applied to naturally occurring microbial systems. The microbial communities of the human microbiome offer a rich opportunity for such study. Viruses have been noted in numerous human body regions, viral communities have been studied

in such diverse locations as the oral cavity, skin, bloodstream, gut, and respiratory tract [2], and direct effects on microbial communities by phage have been observed [72]. In addition to providing a scientifically useful study system, these human-associated communities can have serious impacts on the health of the human host, and understanding microbe-virus interactions could uncover information pertinent to human health.

One such health-linked community is the human vaginal microbiome. The microbial composition of the vagina has been implicated in complications including bacterial vaginosis [70,78], sexually transmitted diseases [60], and risk of preterm birth [44,69,87]. There is no single core vaginal microbiome; rather, a set of five general types has been identified. In four of these types, the microbiome is heavily dominated by a single *Lactobacillus* species (*L. crispatus, L. gasseri, L. iners,* or *L. jensenii*); the fifth is characterized by a consortium of strict anaerobes [71]. Regardless of community type, pregnancy has distinct effect on the vaginal microbiome: the community shifts toward lower diversity, lower anaerobe abundance, and higher *Lactobacillus* abundance [1,76,85]. The relatively low species diversity in the vaginal microbiome during pregnancy provides an opportunity to analyze diversity in the CRISPR system using metagenomic techniques, as sufficient sequencing depth of the dominant *Lactobacillus* species is more practical to achieve. Furthermore, Lactobacillus strains commonly contain CRISPRs; as of the close of 2015, 63% of sequenced *Lactobacillus* representatives contain a CRISPR system [28].

Using vaginal microbiome samples collected from a cohort of pregnant women at low and high risk for preterm birth, I extracted CRISPR spacers and characterized variation within and between individual women. Analyzing these CRISPRs revealed that while variation in spacers present between women was high, persistence of a single CRISPR-type within an individual was common, with limited variation over the course of a pregnancy. Some links between spacers and matched protospacer presence were also observed within one sample.

Microbial colonization also contributes significantly to the health of individuals with cystic fibrosis. Among the most common life-shortening genetic disorders, cystic

fibrosis results from loss of function of the cystic fibrosis transmembrane conductance regulator (CFTR), a chloride ion transporter. This leads to a buildup of thick, sticky mucus in the lungs. In the healthy human lung, the microbiome is transient; new microbes are constantly introduced through inhalation and eliminated through mucociliary clearance [23]. However, in cystic fibrosis patients, the abnormal mucus prevents clearance of microbes and facilitates long-term respiratory infections, which are responsible for the majority of cystic fibrosis morbidity and mortality [21].

Among the most common and damaging of these pathogens is *Pseudomonas aeruginosa*, a ubiquitous opportunistic pathogen. Initial colonization of the lungs is believed to stem from environmental strains encountered by the patient which then adapt to the lung environment [36]. Established *P. aeruginosa* infections are extremely tough to eradicate due to frequent antibiotic resistance and formation of biofilms [24]. Patients are also susceptible to colonization by epidemic *P. aeruginosa* strains, which are especially adept at infecting cystic fibrosis patients and are associated with worse clinical outcomes [26]. Epidemic strains are capable of infecting lungs already colonized by non-epidemic strains, and some of their competitive advantages have been linked to integrated prophages. In the genomes of Liverpool epidemic strains (LES), disrupting some of these prophages put the strain at a competitive disadvantage relative to its intact ancestor in a rat lung chronic infection model [88]. These prophages have also been shown to retain their lytic activity and may affect *P. aeruginosa* density in the cystic fibrosis lung [47]. While LES strains use integrated phages to their advantage, *P. aeruginosa* also harbors a number of antiviral defenses, including a CRISPR system which is known provide immunity against phages [16]. Several *P. aeruginosa* phages are known to produce anti-CRISPR proteins; in fact, it was the species in which anti-CRISPRs were first discovered [10,11,67].

Given the evidence for phage and prophage activity and influence on *P. aeruginosa* in the cystic fibrosis lung microbiome and the potential for *P. aeruginosa* CRISPRs to modulate phage influence, I investigated the extent of interaction with phage by assessing CRISPR spacer content in 14 metagenomic samples from cystic fibrosis patients. In keeping with my analysis of the vaginal microbiome, no diversity was observed in *P. aeruginosa*

within a sample, but CRISPR spacer content was completely different between patients. To further investigate diversity in the global population, I proceeded to assess a large pool of *P. aeruginosa* genomes from varied environments. I found a diverse array of spacers matching a broad variety of sequenced phage. These diverse spacers were used as sequence tags for a novel method of phage tracking called prototyping which I used to examine the ecology of *P. aeruginosa* phage populations, revealing that both hosts and phages have a panmictic population structure.

References

[1] Aagaard, K., Riehle, K., et al. (2012). A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One 7* (6), e36466.

[2] Abeles, S.R. and Pride, D.T. (2014). Molecular bases and role of viruses in the human microbiome. *J. Mol. Biol. 426* (23), 3892–3906.

[3] Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol. 62* , 53–70.

[4] Andersson, A.F. and Banfield, J.F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science 320* (5879), 1047–1050.

[5] Barrangou, R., Fremaux, C., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science 315* (5819), 1709–1712.

[6] Becks, L. and Agrawal, A.F. (2010). Higher rates of sex evolve in spatially heterogeneous environments. *Nature 468* (7320), 89–92.

[7] Bell, G. (1982). *The Masterpiece of Nature: The Evolution and Genetics of Sexuality*. University of California Press, Oakland, CA.

[8] Bohannan, B.J.. and Lenski, R.E. (2000). Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecol. Lett. 3* (4), 362–377.

[9] Bohannan, B.J.M. and Lenski, R.E. (1997). Effect of resource enrichment on a chemostat community of bacteria and bacteriophage. *Ecology 78* (8), 2303–2315.

[10] Bondy-Denomy, J., Garcia, B., et al. (2015). Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature 526* (7571), 136–139.

[11] Bondy-Denomy, J., Pawluk, A., et al. (2013). Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature 493* (7432), 429–432.

[12] Brouns, S.J.J., Jore, M.M., et al. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science 321* (5891), 960–964.

[13] Buckling, A. and Hodgson, D.J. (2007). Short-term rates of parasite evolution predict the evolution of host diversity. *J. Evol. Biol. 20* (5), 1682–1688.

[14] Buckling, A. and Rainey, P.B. (2002). Antagonistic coevolution between a bacterium and a bacteriophage. *Proc. R. Soc. B Biol. Sci. 269* (1494), 931–936.

[15] Buckling, A., Wei, Y., et al. (2006). Antagonistic coevolution with parasites increases the cost of host deleterious mutations. *Proc. R. Soc. B Biol. Sci. 273* (1582), 45–49.

[16] Cady, K.C., Bondy-Denomy, J., et al. (2012). The CRISPR/cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *J. Bacteriol. 194* (21), 5728–5738.

[17] Chakraborty, S., Waise, T.M.Z., et al. (2009). Assessment of the evolutionary origin and possibility of CRISPR-Cas (CASS) mediated RNA interference pathway in *Vibrio cholerae* O395. *In Silico Biol. 9* (4), 245–254.

[18] Childs, L.M., Held, N.L., et al. (2012). Multiscale model of CRISPR-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin. *Evolution 66* (7), 2015–2029.

[19] Clarke, B. (1976). The ecological relationships of host-parasite relationships. In A.E.R. Taylor and R. Muller, eds., *Genetic Aspects of Host-Parasite Relationships*. Blackwell, Oxford, 87–103.

[20] Cui, Y., Li, Y., et al. (2008). Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. *PLoS One 3* (7), e2652.

[21] Cystic Fibrosis Foundation. (2015). *Patient Registry Annual Data Report 2014*.

[22] Deveau, H., Barrangou, R., et al. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol. 190* (4), 1390–1400.

[23] Dickson, R.P. and Huffnagle, G.B. (2015). The lung microbiome: New principles for respiratory bacteriology in health and disease. *PLoS Pathog. 11* (7), e1004923.

[24] Drenkard, E. and Ausubel, F.M. (2002). *Pseudomonas* biofilm formation and antibiotic resistance are linked to phenotypic variation. *Nature 416* (6882), 740–743.

[25] Fabre, L., Zhang, J., et al. (2012). CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS One 7* (5), e36995.

[26] Fothergill, J.L., Walshaw, M.J., et al. (2012). Transmissible strains of *Pseudomonas aeruginosa* in cystic fibrosis lung infections. *Eur. Respir. J. 40* (1), 227–238.

[27] Godde, J.S. and Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol. 62* (6), 718–729.

[28] Grissa, I., Vergnaud, G., et al. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics 8*, 172.

[29] Groenen, P.M., Bunschoten, A.E., et al. (1993). Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol. Microbiol. 10* (5), 1057–1065.

[30] Haerter, J.O. and Sneppen, K. (2012). Spatial structure and Lamarckian adaptation explain extreme genetic diversity at CRISPR locus. *mBio 3* (4), e00126-12.

[31] Haerter, J.O., Trusina, A., et al. (2011). Targeted bacterial immunity buffers phage diversity. *J. Virol. 85* (20), 10554–10560.

[32] Haft, D.H., Selengut, J., et al. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol. 1* (6), e60.

[33] Hale, C.R., Zhao, P., et al. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell 139* (5), 945–56.

[34] Hamilton, W.D. (1982). Pathogens as causes of genetic diversity in their host populations. In R.M. Anderson and R.M. May, eds., *Population Biology of Infectious Diseases*. Springer-Verlag, New York, 269–296.

[35] Hamilton, W.D., Axelrod, R., et al. (1990). Sexual reproduction as an adaptation to resist parasites (a review). *Proc. Natl. Acad. Sci. 87* (9), 3566–3573.

[36] Hauser, A.R., Jain, M., et al. (2011). Clinical significance of microbial infection and adaptation in cystic fibrosis. *Clin. Microbiol. Rev. 24* (1), 29–70.

[37] He, J. and Deem, M.W. (2010). Heterogeneous diversity of spacers within CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats). *Phys. Rev. Lett. 105* (12), 128102.

[38] Held, N.L., Childs, L.M., et al. (2013). CRISPR-Cas systems to probe ecological diversity and host-viral interactions. In *CRISPR*. Springer Press, 221–250.

[39] Held, N.L., Herrera, A., et al. (2010). CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS One 5* (9), e12988.

[40] Held, N.L., Herrera, A., et al. (2013). Reassortment of CRISPR repeat-spacer loci in *Sulfolobus islandicus*. *Environ. Microbiol. 15* (11), 3065–3076.

[41] Hermans, P.W., van Soolingen, D., et al. (1991). Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect. Immun. 59* (8), 2695–2705.

[42] Horvath, P., Coûté-Monvoisin, A.-C., et al. (2009). Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int. J. Food Microbiol. 131* (1), 62–70.

[43] Horvath, P., Romero, D.A., et al. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol. 190* (4), 1401–1412.

[44] Hyman, R.W., Fukushima, M., et al. (2014). Diversity of the vaginal microbiome correlates with preterm birth. *Reprod. Sci. 21* (1), 32–40.

[45] Iranzo, J., Lobkovsky, A.E., et al. (2013). Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-cas in an explicit ecological context. *J. Bacteriol. 195* (17), 3834–3844.

[46] Jaenike, J. (1978). An hypothesis to account for the maintenance of sex within populations. *Evol. Theory 3* , 191–194.

[47] James, C.E., Davies, E.V., et al. (2015). Lytic activity by temperate phages of *Pseudomonas aeruginosa* in long-term cystic fibrosis chronic lung infections. *ISME J. 9* (6), 1391–1398.

[48] Jansen, R., Embden, J.D.A. van, et al. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol. 43* (6), 1565–1575.

[49] Jore, M.M., Brouns, S.J.J., et al. (2012). RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. *Cold Spring Harb. Perspect. Biol. 4* (6), a003657.

[50] Kamerbeek, J., Schouls, L., et al. (1997). Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol. 35* (4), 907–914.

[51] Koonin, E.V. and Wolf, Y.I. (2009). Is evolution Darwinian or/and Lamarckian? *Biol. Direct 4* (1), 42.

[52] Koskella, B. and Lively, C.M. (2009). Evidence for negative frequency-dependent selection during experimental coevolution of a freshwater snail and a sterilizing trematode. *Evol. Int. J. Org. Evol. 63* (9), 2213–2221.

[53] Labrie, S.J., Samson, J.E., et al. (2010). Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol. 8* (5), 317–327.

[54] Lenski, R.E. (1988). Experimental studies of pleiotropy and epistasis in *Escherichia coli*. I. Variation in competitive fitness among mutants resistant to virus T4. *Evolution 42* (3), 425–432.

[55] Levy, A., Goren, M.G., et al. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature 520* (7548), 505–510.

[56] Lotka, A.J. (1920). Undamped oscillations derived from the law of mass action. *J. Am. Chem. Soc. 42* (8), 1595–1599.

[57] Makarova, K.S. and Koonin, E.V. (2013). Evolution and Classification of CRISPR-Cas Systems and Cas Protein Families. In R. Barrangou and J. van der Oost, eds., *CRISPR-Cas Systems*. Springer Berlin Heidelberg, 61–91.

[58] Makarova, K.S., Wolf, Y.I., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol. 13* (11), 722–736.

[59] Marraffini, L.A. (2015). CRISPR-Cas immunity in prokaryotes. *Nature 526* (7571), 55–61.

[60] Mastromarino, P., Di Pietro, M., et al. (2014). Effects of vaginal lactobacilli in *Chlamydia trachomatis* infection. *Int. J. Med. Microbiol. 304* (5–6), 654–661.

[61] Modi, S.R., Lee, H.H., et al. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature 499* (7457), 219–222.

[62] Mojica, F.J., Ferrer, C., et al. (1995). Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. *Mol. Microbiol. 17* (1), 85–93.

[63] Mojica, F.J.M., Díez-Villaseñor, C., et al. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol. 60* (2), 174–182.

[64] Mojica, F.J.M., Díez-Villaseñor, C., et al. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology 155* (3), 733–740.

[65] Nee, S. (1989). Antagonistic co-evolution and the evolution of genotypic randomization. *J. Theor. Biol. 140* (4), 499–518.

[66] Otto, S.P. and Nuismer, S.L. (2004). Species interactions and the evolution of sex. *Science 304* (5673), 1018–1020.

[67] Pawluk, A., Bondy-Denomy, J., et al. (2014). A new group of phage anti-CRISPR genes inhibits the Type I-E CRISPR-Cas system of *Pseudomonas aeruginosa*. *mBio 5* (2), e00896-14.

[68] Peters, A.D. and Lively, C.M. (1999). The Red Queen and fluctuating epistasis: A population genetic analysis of antagonistic coevolution. *Am. Nat. 154* (4), 393–405.

[69] Petricevic, L., Domig, K.J., et al. (2014). Characterisation of the vaginal *Lactobacillus* microbiota associated with preterm delivery. *Sci. Rep. 4* , 5136.

[70] Ravel, J., Brotman, R.M., et al. (2013). Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. *Microbiome 1* (1), 29.

[71] Ravel, J., Gajer, P., et al. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U. S. A. 108 Suppl 1* , 4680–4687.

[72] Reyes, A., Wu, M., et al. (2013). Gnotobiotic mouse model of phage–bacterial host dynamics in the human gut. *Proc. Natl. Acad. Sci. 110* (50), 20236–20241.

[73] Riehm, J.M., Vergnaud, G., et al. (2012). *Yersinia pestis* lineages in Mongolia. *PLoS One 7* (2), e30624.

[74] Rodriguez-Brito, B., Li, L., et al. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J. 4* (6), 739–751.

[75] Rohwer, F. and Thurber, R.V. (2009). Viruses manipulate the marine environment. *Nature 459* (7244), 207–212.

[76] Romero, R., Hassan, S.S., et al. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome 2* (1), 4.

[77] Shariat, N., DiMarzio, M.J., et al. (2013). The combination of CRISPR-MVLST and PFGE provides increased discriminatory power for differentiating human clinical isolates of *Salmonella enterica* subsp. *enterica* serovar Enteritidis. *Food Microbiol. 34* (1), 164–173.

[78] Srinivasan, S., Hoffman, N.G., et al. (2012). Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS One 7* (6), e37818.

[79] Sun, C.L., Barrangou, R., et al. (2013). Phage mutations in response to CRISPR diversification in a bacterial population. *Environ. Microbiol. 15* (2), 463–470.

[80] Takeuchi, N., Wolf, Y.I., et al. (2012). Nature and intensity of selection pressure on CRISPR-associated genes. *J. Bacteriol. 194* (5), 1216–1225.

[81] Touchon, M., Charpentier, S., et al. (2011). CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *J. Bacteriol. 193* (10), 2460–2467.

[82] Touchon, M. and Rocha, E.P.C. (2010). The small, slow and specialized CRISPR and Anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE 5* (6), e11126.

[83] Tyson, G.W. and Banfield, J.F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol. 10* (1), 200–207.

[84] Van Valen, L. (1973). A new evolutionary law. *Evol. Theory 1* (1), 1–30.

[85] Walther-António, M.R.S., Jeraldo, P., et al. (2014). Pregnancy's stronghold on the vaginal microbiome. *PLoS One 9* (6), e98514.

[86] Weitz, J.S., Hartman, H., et al. (2005). Coevolutionary arms races between bacteria and bacteriophage. *Proc. Natl. Acad. Sci. U. S. A. 102* (27), 9535–9540.

[87] Wen, A., Srinivasan, U., et al. (2014). Selected vaginal bacteria and risk of preterm birth: an ecological perspective. *J. Infect. Dis. 209* (7), 1087–1094.

[88] Winstanley, C., Langille, M.G.I., et al. (2009). Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res. 19* (1), 12–23.

[89] Winter, C., Bouvier, T., et al. (2010). Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "Killing the Winner" hypothesis revisited. *Microbiol. Mol. Biol. Rev. 74* (1), 42–57.

CHAPTER 2

CRISPR-Induced Distributed Immunity in Simulated Microbial Populations[1]

Abstract

In bacteria and archaea, viruses are the primary infectious agents, acting as virulent, often deadly pathogens. A form of adaptive immune defense known as CRISPR-Cas enables microbial cells to acquire immunity to viral pathogens by recognizing specific sequences encoded in viral genomes. The unique biology of this system results in evolutionary dynamics of host and viral diversity that cannot be fully explained by the traditional models used to describe microbe-virus coevolutionary dynamics. Here, we show how the CRISPR-mediated adaptive immune response of hosts to invading viruses facilitates the emergence of an evolutionary mode we call distributed immunity - the coexistence of multiple, equally-fit immune alleles among individuals in a microbial population. We use an eco-evolutionary modeling framework to quantify distributed immunity and demonstrate how it emerges and fluctuates in multi-strain communities of hosts and viruses as a consequence of CRISPR-induced coevolution under conditions of low viral mutation and high relative numbers of viral protospacers. We demonstrate that distributed immunity promotes sustained diversity and stability in host communities and decreased viral population density that can lead to viral extinction. We analyze sequence diversity of experimentally coevolving populations of *Streptococcus thermophilus* and their viruses where CRISPR-Cas is active, and find the rapid emergence of distributed immunity in the host population, demonstrating the importance of this emergent phenomenon in evolving microbial communities.

Introduction

All organisms are susceptible to infection by viral pathogens. The sheer number of viruses found in natural environments is staggering; it is estimated that 10^{31} virus particles are circulating at any time [10,54], containing at least hundreds of thousands of

[1] This chapter was originally published as Childs, L.M.*, England, W.E.*, Young, M.J., Weitz, J.S., and Whitaker, R.J. (2014). CRISPR-Induced Distributed Immunity in Microbial Populations. PLoS ONE *9*, e101710, and is reprinted here with permission.

genotypes [3], most of which infect bacteria and archaea. Bacteria and archaea resist infection through random mutation, resulting in loss or modification of viral receptors, or through targeted defense systems such as physical blocking, restriction-modification systems, and abortive infection systems [8,15,33,36,47,56,63]. Both negative frequency-dependent selection (NFDS) and diversifying selection for microbial resistance have been suggested to result in the diversity observed in natural systems [4,40,50]. The trade-off between resistance and growth rate has become the dominant model for microbe-virus coevolution [64], with variation in fitness driving diversification of the host and resulting in the predicted coexistence of many genotypes of both hosts and viruses [61]. These theoretically predicted trade-offs have also been seen to promote diversity of both host and viral populations in experimentally evolved populations [9,12,13,22,37,42].

Recently the CRISPR-Cas system was experimentally shown to function as an adaptive microbial resistance mechanism, using the model organism *Streptococcus thermophilus* [5] (see reviews in [1,6,8,18,19,31,32,44,48,55,62]). The CRISPR-Cas system, components of which are found in the majority of sequenced microbes [25], is comprised of short DNA fragments (spacers) flanked by palindromic repeats in repeat-spacer arrays [1]. These fragments are often identical to sequences in plasmids, viruses, and other foreign elements [45]. When a microbe containing an active CRISPR system encounters one of these foreign elements, it can add a new spacer matching a sequence in the foreign genome (protospacer) [5]. The CRISPR system can acquire spacers from many locations in a foreign genome, requiring only a short protospacer-associated motif (PAM) adjacent to the protospacer [18,46]. Repeat-spacer arrays are transcribed, processed, and used to guide an effector complex which inactivates matched foreign genetic material on any subsequent encounter [11]. Escape mutations in protospacers prevent recognition by the CRISPR-Cas system resulting in a coevolutionary dynamic in which viruses evolve through random mutation while hosts evolve through "directed mutation" facilitated by adaptive immunity [18,41,52,53].

We propose that crucial elements of the CRISPR system result in a diversifying coevolutionary mode that is distinct from the traditional trade-off model described above. Adaptive CRISPR acquisition of new spacers leads to the potential for competing

CRISPR genotypes to emerge within a host population at the same (or similar) time – akin to the phenomena of "clonal interference" [17,24]. The vast reservoir of protospacers in each virus creates the potential for competing host genotypes with similar (or identical) resistance phenotypes that are not necessarily subject to fitness tradeoffs between immune alleles. In contrast, viral strains are limited in potential escape mutations by fitness constraints on mutations in their genomes that can modify regulatory elements and RNA- and protein-encoding genes. In addition, each viral escape mutant only allows access to a single host immune allele, potentially composing a small subset of the host population [21,38]. We hypothesized that these differences would allow for a dramatic restructuring of the coevolutionary mode wherein many different hosts are immune to the same virus in different ways. We label this many-to-one, genotype-to-phenotype phenomenon *distributed immunity*.

We previously developed an eco-evolutionary model of CRISPR-mediated host-viral coevolution [14]. In brief, the model incorporates density-dependent Lotka-Volterra like ecological dynamics with the evolutionary introduction of new hosts and viral strains with novel genetic states. Ecological rules of interaction including host reproduction and death, viral infection of hosts and viral deactivation outside of hosts determine host and viral densities. Viral infection of hosts can lead to either host lysis or viral deactivation, which may occur with or without spacer integration. During replication, viral strains evolve through mutation generating a novel protospacer. Host immunity is determined by the presence of at least one spacer matching a virus, yet is not foolproof, i.e., there is a small chance that a host with a matching spacer to an infecting virus will not be immune [14]. In simulations of our model, host and viral populations oscillate in abundance over short time scales, whereas host and viral genotype composition changes over long time scales, mediated by coevolutionary adaptation. A comparison of this and other models of CRISPR-mediated coevolutionary dynamics (e.g., [26,27,34,39,59]), whose exact dynamics depend on the specific molecular, ecological and evolutionary parameters can be found elsewhere [14,28].

Within our model, examining the diversity of the host population at each maximum in total host population abundance (host peaks), we observed two types of emergent

population dynamics: (i) near selective sweeps by novel or recurring strains and (ii) simultaneous growth of phenotypically similar but genotypically diverse groups of strains which we termed coalitions [14]. Although the diversification of host populations with CRISPR immunity had been noted previously [14,27,34,43,59], in this paper, we present a metric, *population-wide distributed immunity* (PDI) to quantify distributed immunity in a population, to examine how distributed immunity varies over time and to determine how this evolutionary mode affects the coevolutionary dynamic. We used simulated data from our model to: (i) determine when coalitions are characterized by distributed immunity; (ii) identify conditions under which distributed immunity is the dominant evolutionary mode in a simulation; and (iii) quantify the effects of distributed immunity on host-viral relationships by examining diversity and stability of host and viral populations. Finally we determined that the diversity exhibited in an experimental host-viral community is associated with distributed immunity.

Results

Quantifying distributed immunity

Distributed immunity denotes the emergent phenomenon in which multiple immune alleles coexist within and between hosts. When these alleles are distributed between different hosts that have CRISPR-Cas resistance, then multiple hosts have similar immune phenotypes yet have distinct, coexisting associated CRISPR genotypes. To measure the impact of distributed immunity, on each population, we developed a metric called population-wide distributed immunity (PDI) in which CRISPR-Cas immune relationships of all host-host-viral strain triplets are tested to determine if the two host strains contain spacers matching different protospacers on the same viral strain (Figure 2.1, see Methods for details of the calculation). The intuition behind our metric is that all triplets contribute positively to PDI when both hosts are immune to the virus by means of distinct spacers matching the virus. In the case where both hosts are immune to the virus but via the identical spacer, the immunity is not distributed throughout the population and thus does not contribute to PDI. Although phenotypically immunity via identical or distinct spacers is equivalent, the varied genotypes may follow different evolutionary pathways. For example, when PDI is high, mutation of a single protospacer does not

permit escape in the majority of the host population. However, when PDI is low, a single protospacer mutation may lead to viral escape in most of the host population. The degree of contribution by each triplet depends on the product of the relative abundance of the host strains and viral strain and immunity between the host and viral strains (see Methods for details of the calculation). The maximum PDI for a population at any time increases with the number of host strains (with n host strains the maximum is 1-1/n) and is only obtainable when the following hold: there are at least two alleles that confer immunity to the viral strains, all host strains are immune to viral strains, and the abundance of each host CRISPR allele is equal (Figure 2.2). Note that the abundance of the viral strains does not affect the potential for PDI (see SI text for further discussion).

In the simulated eco-evolutionary dynamics of hosts and viruses [14], we find that PDI varies through time (Figure 2.3). PDI is typically highest just prior to peaks in host population density and drops to at or near zero in between (Figure 2.3). Every peak of host density does not contain high PDI, even if its potential maximum PDI is high, and in our simulations we find that measured PDI is well below the potential maximum. Low PDI results from (i) unevenness of the host population (Figure 2.3b-1, 2.2), (ii) a large fraction of the hosts lacking immunity to the viral population (Figure 2.3b-2) or (iii) the majority of hosts having immunity to the viral population via the same spacer (Figure 2.3b-4). In contrast, high PDI occurs when multiple hosts have unique spacers to the same viral strains. This can occur when a dominant host strain diversifies via the acquisition of unique spacers to the same viral strain (Figure 2.3b-3). Across all simulations, the PDI at host peaks ranges from 0 to 0.7203 with an overall mean of 0.0710. We find no direct, predictable relationship between the abundance of host and viral populations at their peaks in relation to the concurrent value of PDI within a single simulation. In contrast, we hypothesize that PDI functions to alter the future host and viral dynamics within a community. Diversified hosts (with a high PDI) may affect the composition and total density of virus populations that recur in the next peak in host density or much later. This is due to the complexity and diversity of both host and viral populations in which a particular diversified host can be targeted by divergent low abundance viruses that were created much earlier.

Parameters that increase population-wide distributed immunity

To determine how biological parameters might influence the evolutionary mode across a simulation toward or away from distributed immunity, we altered four parameters that vary between microbial and viral strains: viral mutation rate, μ; spacer acquisition rate, q; maximum host spacer number, S; and viral protospacer number, P. To avoid the period of transient dynamics occurring at the initiation of the simulations from a single viral and single host strain, we measure median PDI in the last 500 hours of each simulation, where the host spacer locus is filled and both host and viral diversity are most regular (see SI text, Figure 2.4). Comparing the population dynamics between sets of simulations with varying parameters, we found that average PDI across the simulations increases when viral mutation rate decreases and when the number of relative protospacers increases (Figure 2.5). There are also increases in PDI when the spacer acquisition rate increases and the number of spacers increases, but PDI above 0.1 is rarely seen (Figure 2.5). The highest average PDI is seen with high relative protospacer number (P=20) and low viral mutation rate ($\mu = 10^{-7}$) while lowest average PDI occurs with low relative protospacer number (P=5) and low spacer acquisition rate (q=10^{-6}). Increases in average PDI result from coevolutionary dynamics that include more host population peaks with higher PDI, rather than from an increase in PDI when host populations are not near their peak values.

Population-wide distributed immunity is associated with individual distributed immunity

In simulations with a higher average PDI, we observed an additional dynamic where individual host genotypes contain multiple spacers matching the same viral strain at distinct protospacers. This represents an analogous form of distributed immunity, albeit within a single host. Since this will have similar evolutionary effects as PDI, we quantify the average per host immunity to viral strains with a new metric denoted as individual distributed immunity (IDI). IDI is equal to the average number of distinct matching spacers between each pair of viral and host strains (see Methods for details of the calculation). When IDI is greater than one, the host population is on average immune in multiple ways to the viral population due to targeting multiple regions of the viral genome. We find that there is strong correlation between PDI and IDI (Figure 2.6) and, as

with PDI, there is high IDI with low viral mutation rate and high protospacer number (Figure 2.7). Hereafter, we collectively refer to PDI and IDI as DI.

Elevated distributed immunity is associated with increases in host diversity, density, and stability

Having identified conditions under which simulations with high levels of distributed immunity are linked to changes in host-virus relationships, we investigated possible consequences of these altered interactions. We found that simulations resulting in high levels of distributed immunity are correlated with increased host strain count and population density (Figure 2.8A-D). We find a much stronger association between DI and these population level indicators than when evaluating the statistical relationship between mutation rate and protospacer number alone. For example, the Spearman rank correlation coefficient between host population density and PDI is 0.84 whereas it is -0.31 and 0.49, when evaluated against mutation rate and *P*, respectively (all $p<0.001$). Similarly, the Spearman rank correlation coefficient between host strain count and PDI is 0.78 whereas it is -0.26 and 0.27, when evaluated against mutation rate and *P*, respectively (all $p<0.001$). The data collapse of host population density and host strain count as a function of PDI from simulations with different governing parameters is apparent in Figure 2.8A-D. Investigating simulations where distributed immunity has a strong effect (high DI), we also observed extended periods of high density, stable host populations (see time points between 9700-10000 in Figure 2.9A-C for a typical example). Periods of stable host-controlled dynamics occur exclusively in parameter sets which have higher DI: $P=15$, $P=20$, and $\mu=10^{-7}$, and the proportion of simulations which exhibit extended stable periods increases with increasing DI (Figure 2.9E, black bars). The finding of extended stability is not driven solely by the extended high host density; this pattern is observed whether DI is measured at all time points (as in Figure 2.9E), or only at host density peaks.

Elevated PDI is associated with decreased viral diversity and density

In contrast to the increases in host population density and host strain count as PDI increases, the trends for viral population density and viral strain count are non-monotonic

(Figure 2.8E-H and S5). At lower PDI (PDI < 0.2) increases in PDI correlate with increases in viral population density and weakly correlate with increases in viral strain count (Figure 2.8 and 2.10). The observed viral population increases are also correlated with increases in host population size and host immunity (Figure 2.11). Although immunity is increasing, it is still relatively low, suggesting that individual viral strains can continue to grow on subsets of the total host population. Simultaneously, as PDI increases, the host population is also increasing, so that each subset of hosts that viruses can infect is actually larger than at lower PDI. At higher PDI (PDI > 0.2), increases in PDI correlate with decreases in viral population density and viral strain count (Figure 2.8). Beyond PDI = 0.2, increases in host population size and immunity no longer correspond to higher viral densities. This decrease in viral density is consistent with the fact that the proportion of hosts that viruses can infect (HVI, see Methods for details of the calculation) decreases as DI increases, and HVI is significantly lower in simulations with higher DI (Figure 2.12). Accompanying decreases in viral population sizes we find that the proportion of simulations in which viruses go extinct increases with increasing DI (Table 2.1 and Figure 2.13, dark gray bars). Parameter sets with the highest DI, P=20 and $\mu = 10^{-7}$, result in viral extinction in 10% and 12% of simulations with filled loci, respectively, the highest rates of extinction of any parameter set (Table 2.1). Considering simulations in which the CRISPR locus does not fill before the last 500 hours, 90.7% end in viral extinction, including 94.3% and 91.6% of P=20 and $\mu = 10^{-7}$ simulations, respectively. Nearly all simulations with lower DI reach a full spacer locus prior to the final 500 hours (Table 2.1).

Elevated distributed immunity identified in an experimental viral-host community

We examined whether the dynamic of distributed immunity observed in simulations is consistent with patterns observed in experimental microbial communities in which both virus and host sequence is known. To do so, we estimated DI within an experimental set of host and viral populations. A quantitative assessment of the contribution of the relative DI to the maintenance of diversity in natural microbial populations is not possible in most studies, as the contemporary virus population is not typically sequenced. Despite technical challenges to date in testing distributed immunity in natural populations, studies

in laboratory populations offer an opportunity to measure distributed immunity. Numerous studies in laboratory populations have shown that upon challenge by a single phage, multiple *S. thermophilus* genotypes emerge with different spacers providing immunity [5,18,32,41,49,53]. For our analysis, we used data from Sun *et al.* [53], the only study with both sequences and abundances from the entire coevolving host and viral populations as required to measure DI. In this study, a laboratory-coevolved population of *Streptococcus thermophilus* and its phage 2972 was found to exhibit rapid spacer addition as well as phage CRISPR escape mutations. After 1 week of co-culture, the host had added 43 new spacers to one CRISPR locus, and three viral mutations in targeted protospacers or PAMs were detected [53]. Given the diversity of new spacers matching a small pool of viral types, we estimated a high value of PDI for these populations. Using populations reconstructed from spacer-containing reads and viral SNP distributions (Figure 2.13, see Methods), the value of PDI after 1 week of coevolution was 0.4331, out of a maximum possible PDI of 0.5933. This estimate of elevated PDI complements Sun *et al*.'s [53] observation of multiple acquisitions of distinct CRISPR escape mutants, and suggests a population-level effect that may act synergistically with individual host-viral interactions. Note that this PDI value is larger than the median PDI in 99.8% and the highest observed PDI in 75.9% of all simulations we conducted. The value of IDI, 1.2264, was higher than the median IDI in 97.7% and the highest observed IDI in 58.7% of simulations.

Discussion

We have explored the immune dynamics resulting from a computational eco-evolutionary model driven by CRISPR-mediated immunity. The model demonstrates how a host-viral community can evolve a complex structure where different hosts are immune to the same virus as a result of immunity conferred by different immune alleles, which we have quantified as distributed immunity. Immunity relationships between hosts and viruses with distributed immunity may appear similar from the phenotype level to relationships lacking distributed immunity; however, the underlying genetic diversity present in distributed immunity changes the dynamics of coevolution. In particular, during periods of elevated distributed immunity, the host population is diverse and stable while the viral

population is restricted in the number and extent of possible beneficial mutations and is prone to extinction. The stable maintenance of multiple non-dominant genotypes that accompanies distributed immunity is likely facilitated by NFDS. The generation of distributed immunity and the selective mechanisms of NFDS may work together to promote diversity.

Several CRISPR models have previously observed diversity in host spacer content both at an individual and population level [14,27,34,59], but understanding that diversity has been a recent exploration. Although Iranzo *et al.* [34] established several population-level findings, such as CRISPR immunity promoting the coexistence of viruses and hosts at intermediate viral mutation rate and the lack of increased viral diversity with CRISPR immunity, they did not attempt to expound upon these findings, which they labeled counterintuitive. Our model, even with its reduced complexity as we ignore populations lacking CRISPRs, is able to reproduce these results and offer an explanation for them via distributed immunity. Here, we have demonstrated that the consequences of viral protospacer number and mutation rate as well as host spacer acquisition rate and spacer number on the population dynamics can be explained as acting through distributed immunity thereby linking the molecular and evolutionary mechanisms to the eco-evolutionary dynamics that have been observed. Since distributed immunity only requires some of the spacers to be distinct, it is consistent with a previously posed model where random deletion lead to selective sweeps of trailer-end spacers [59].

CRISPR-Cas diversity varies greatly among systems. At one end of the spectrum are the slowly-evolving CRISPR-Cas systems of *Escherichia* and *Salmonella*, where estimates indicate that strains that have diverged in the last thousand years have identical CRISPR loci [57]. At the other end are natural populations exhibiting high CRISPR-Cas diversity, including the human gut microbiome [51], *Yersinia pestis* plague foci [16], and hot spring populations of *Sulfolobus islandicus* [29,30]. Notably, in the case of *S. islandicus*, these archaeal populations do not contain a dominant genotype or display evidence of selective sweeps over a ten-year interval [30] but maintain diversity at both the leader and trailer ends of the CRISPR loci over time. Some natural populations demonstrate evidence of past selective sweeps in the form of conserved trailer-end spacers, particularly

populations of acidophilic microbes found in acid mine drainage [2,58,59]. The difference between the immune structures of different microbial populations may be driven by differences in the extent of distributed immunity within populations, differences in the levels of reassortment of CRISPR alleles between strains in different populations [30], or the action of other host defense systems operating along with CRISPR-Cas immunity.

Indeed, our model suggests that the biology of CRISPR-Cas system might define the resulting level of diversity observed in natural populations. We show that the number of protospacers, viral mutation rate, and host acquisition rate all significantly influence the level of distributed immunity in a way that would result in different immune structures in natural populations. These factors have been shown to vary in natural microbial populations. For example, in microbes with active CRISPR-Cas defense, the number of protospacers is determined by both the length of the viral genome and the length and sequence of the PAM sequences, which direct acquisition and interference. We infer that protospacer number is positively correlated with distributed immunity because at higher protospacer numbers it is easier for hosts to acquire multiple spacers to the same virus (higher IDI) and for different hosts to acquire different spacers (higher PDI). We hypothesize that microbial hosts utilizing shorter PAMs or that are infected by viruses with larger genomes are more likely to display a diversified immune structure that is consistent with distributed immunity. Variation in viral mutations rates has also been observed in natural populations. For example, it has been suggested that thermophiles and their viruses have lower mutation rates than their mesophilic counterparts [20,60]. Our model suggests that this is consistent with data showing that the thermophilic archaeon *S. islandicus* appears to maintain a stable diversified population over time [29,30]; however, this hypothesis must be explicitly tested. Finally, in this study we did not explore variation in the probability that CRISPR immunity fails such that a host cell does not recognize and clear a virus for which it has a matching spacer. Such failure may result in the proliferation of a virus to which there exists some immunity in the population. Given our previous analysis showing the relatively minor effects of such failure on resulting dynamics [14] , we do not expect significant effects of the stochastic failure of host spacers on distributed immunity, at least in the range of failure values

observed experimentally [5]. However, in the case of exposure to plasmids rather than viruses, such failure may permit the exchange of genetic material between hosts [35]. Under conditions when genetic exchange is advantageous (e.g., in the presence of many beneficial plasmids [23]) then the occurrence of distributed immunity may result, even if seemingly unfavorable, to protect against virulent viruses.

Although natural population data is not yet available to employ our novel metrics PDI and IDI for quantifying distributed immunity, we have quantified this evolutionary mode in an experimental population. Qualitatively, Sun *et al.* [53] observed rapid transition from clonal to diversified in both host and viral populations as a result of CRISPR-Cas immunity. We demonstrated that this diversification also exhibited rapid emergence of DI and hypothesize that our finding of highly elevated PDI in Sun *et al.* [53] may be due, in part, to the relatively large number of protospacers in the genomes of phage (associated with replete PAMs), as compared to the use of low number of protospacers (P = 5-20) in our models due to computational constraints. This hypothesis is further supported by our simulation results where DI increases as we increase protospacer number (see Figure 2.5A). We predict that in this system when the *S. thermophilus* hosts exhibit distributed immunity, viral populations will be smaller, less diverse and more prone to extinction. We consider it an important future goal to extend the DI analysis of *S. thermophilus* and phage to systems in which host and viral metagenomes are available to further quantify the variation of DI in natural populations.

A better understanding of CRISPR-mediated coevolutionary dynamics will have important implications for medical applications for example those seeking to target microbial pathogens with phage therapy. In addition, our model suggests possible optimal strategies for engineering stable microbial communities immune to phage attack such as those used in biofuels production or other industrial applications. Finally, CRISPR immunity serves as an interesting model system in which to study the broader effects of diversified immunity on pathogen evolution. Such diversity impacts the trajectory of host-virus coevolution in microbes mediated by CRISPR-Cas immunity. Further understanding how distributed immunity affects the evolutionary path of

populations may yield insight into the effects of host immune diversity in microbial communities and other systems.

Methods

Model information and statistical analyses

We use the model introduced in Childs *et al.* [14] to generate our simulation data. Briefly, in the model, ecological host-viral dynamics are combined with the introduction of new host and viral strains through changes in the CRISPR space and protospacer states. Hosts may acquire new spacers during viral infection, and viruses may mutate to novel protospacers during replication. Host immunity towards an infecting virus requires the presence of at least one spacer matching a viral protospacer, but is not full proof. The population dynamics of host and viral strains are deterministic but the incorporation of hosts' spacers and mutation of viral protospacers occurs stochastically. Further details of the model are reviewed in the supplemental information with the parameters used in Table 2.2. Although this paper focuses on four parameters (protospacer number, spacer number, viral mutation rate, spacer acquisition rate), Childs *et al.* [14] more thoroughly tests dependencies of model dynamics on other parameters. Due to the stochastic nature of our model, the parameter regions surveyed were limited by computational cost. All results presented are averages of 200 replicate simulations, unless otherwise noted (Table 2.1), with each replicate represented by the median value across the final 500 hours of that simulation. One hour is equivalent to the inverse of the growth rate – what we denote here as a typical host generation time. Simulations were excluded from population averages whenever the spacer states did not contain the maximum number of spacers (full locus) throughout the final 500 hours of simulation or whenever the viral population fell below our density cutoff before the locus was filled (Table 2.1).

For each of the four parameters varied (protospacer number, spacer number, viral mutation rate, spacer acquisition rate), measurements from replicates at each parameter value tested were grouped. The means of replicate PDI and IDI measurements were compared using analysis of variation for unbalanced data (data from Figures 2.5 and 2.7).

The Spearman rank correlation coefficients were determined for variations in each parameter between PDI, host population density, viral population density, host strain count, viral strain count, and IDI (data from Figures 2.5, 2.6 and 2.7). The Spearman rank correlation coefficients were also determined for variations in PDI, host population density and immunity combining all parameter sets (data from Figure 2.11). R^2 values were determined for correlations between HVI and PDI, and between HVI and IDI (data from Figure 2.12).

The data collapse of host and viral output variables, as a function of PDI, from simulations with different governing parameters is apparent in Figures 4 and S5. To test for correlations, linear R^2 values were determined for variations in each parameter between PDI, host population density, viral population density, host strain count, viral strain count and IDI for variations in each parameter (data from Figure 2.8, 2.10). Despite significant linear correlation in almost all cases, except between PDI, host strain count and viral strain count when varying S, it was evident upon inspection that the relationships between PDI and viral population density and viral strain count were better described by non-linear functions, particularly quadratic functions. To quantify this, we fit a quadratic model for viral output parameters and compared the quality of fit to a linear model using AIC; the relationship of all PDI and viral output statistics were better fits as demonstrated by lower AIC values except for PDI and viral strain count when varying S where both linear and quadratic fits were not significant (see Table 2.3).

To compare the proportion of simulations that are stable, fluctuating, or end in viral extinction, 10,000 random subsamples of 230 simulations (10% of the total simulations with filled loci) were taken. The mean proportions of simulations in each bin that fell into the stable or viral extinction category were compared using analysis of variation (data from Figure 2.13). We define a population to be stable when the host population exceeds 3e5 for more than 100 hours (approximately 95% of the carrying capacity).

Population-wide distributed immunity (PDI)

To quantify the population-level distribution of immune alleles between hosts with similar immune phenotypes but distinct CRISPR genotypes, we compare all triplets of

two host strains and a viral strain. We determine which triplets contain distinct spacers matching protospacers in the virus to quantify PDI as follows:

$$PDI = \sum_i \sum_j \sum_k \left(1 - \frac{|N_i - N_j|}{\max(N)}\right) \sigma_{ijk} N_i N_j V_k$$

$$\sigma_{ijk} = \begin{cases} 1, & if \quad R(G_i, H_k) R(G_j, H_k) > R^2(G_i, G_j, H_k) \\ 0, & otherwise \end{cases}$$

where N_i is the population proportion of the i^{th} host strain, V_k is the population proportion of the k^{th} viral strain, G_i is the set of spacers belonging to the i^{th} host strain, H_k is the set of protospacers belonging to the k^{th} viral strain, $R(G_i, H_k)$ determines the number of matching spacers and protospacers between the states G_i and H_k, and $R(G_i, G_j, H_k)$ determines the number of matching spacers and protospacers between all the states G_i, G_j and H_k. Further, max(N) denotes the maximum proportion of any given host strain in the population.

Triplets with matching spacers and protospacers contribute to PDI via the function σ. The relative of abundance of the strains from a triplet determines the level of contribution of that triplet to PDI. The total value of PDI is weighted by host strains at or similar to the size of the dominant host strain in order to minimize the summed contribution of numerous strains found at low proportion.

Individual distributed immunity (IDI)

We introduce individual distributed immunity to quantify the distribution of immunity within hosts, in contrast to PDI, which quantifies the distribution of immunity between hosts. IDI is the average number of spacers per host matching the viral population:

$$IDI = \sum_i \sum_k N_i V_k R(G_i, H_k)$$

where the host proportion (N_i), the viral proportion (V_k), the host spacer state (G_i), the viral spacer state (H_k), and the number of matches between spacer and protospacer states $R(G_i,H_k)$ are defined as in PDI.

Hosts that Viruses can Infect (HVI)

The average proportion of hosts that viruses can infect is quantified by HVI:

$$HVI = \sum_i \sum_k N_i V_k \left(1 - M(G_i, H_k)\right)$$

where $M(G_i, H_k)$ determines the presence or absence of matching spacers and protospacers between the states G_i and H_k. The host proportion (N_i), the viral proportion (V_k), the host spacer state (G_i), the viral spacer state (H_k) are defined as in PDI.

Experimental population DI calculations

Sequencing reads from the Sun *et al.* study [53] (accession number SRA049615) containing at least two novel spacers, or at least one novel spacer plus ancestral spacers or leader sequence were considered. Reads were grouped by spacer content; where trailer-end sequence information was not available, the locus was assumed to have the same trailer-end spacers as other reads with similar leader-end spacer content (Figure 2.14). If trailer end spacers could not be inferred in this way, the trailer end was assumed to contain only spacers fixed in the population (Figure 2.14). Each unique set of spacers was considered a host strain; the proportion of reads matching each strain was used for the proportion of each strain in the population (N_i and N_j) for calculation of PDI and IDI. Assuming similar CRISPR loci whenever possible maximizes the number of reads grouped into each CRISPR-type and prevents overestimation of PDI.

Frequencies of three phage mutations in protospacers or PAMs identified by Sun *et al.* [53] were confirmed using breseq [7](available online at http://barricklab.org/breseq). Each possible combination of SNPs was considered a different viral strain. To determine the proportion of phages with each combination of SNPs (SNP-i only, SNP-i and SNP-ii, SNP-i and SNP-iii, or all three SNPs), each mutation was considered an independent event and the probability of each combination was calculated. These proportions were used for V_k in the PDI and IDI equations. Otherwise, PDI and IDI were calculated as in simulated populations.

Figures and Tables

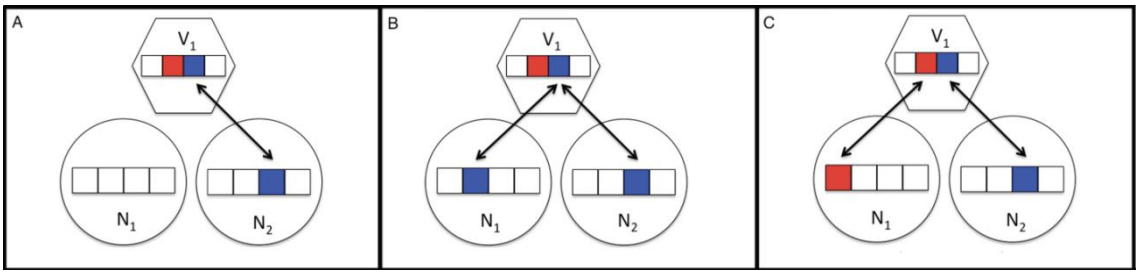


Figure 2.1. Population distributed immunity (PDI) depends on immunity relationships between hosts (circles and viruses (hexagons). Immune elements are denoted as linear arrays of boxes. PDI is the sum of contributions (δPDI) calculated amongst triplets of two hosts and one virus, adjusted by their population proportions, as follows: (A) δPDI = 0 when only one (or neither) hosts in a triplet match the virus as $R(N_1,V_1)=M(N_1,V_1)=0$, $R(N_1,V_1)=M(N_1,V_1)=1$,and $R(N_1,N_2,V_1)=0$. (B) δPDI = 0 when both hosts match the virus with the same spacer as $R(N_1,V_1)=M(N_1,V_1)=1$, $R(N_1,V_1)=M(N_1,V_1)=1$,and $R(N_1,N_2,V_1)=0$. (C) $\delta PDI = N_1N_2V_1\{1-[|N_1-N_2|/\max(N_1,N_2)]\}$ when both hosts match the virus via different spacers as $R(N_1,V_1)=M(N_1,V_1)=1$, $R(N_1,V_1)=M(N_1,V_1)=1$,and $R(N_1,N_2,V_1)=1$. Identical colors, indicated by arrows, represent matching spacer-protospacer pairs. White protospacers and spacers are unique.

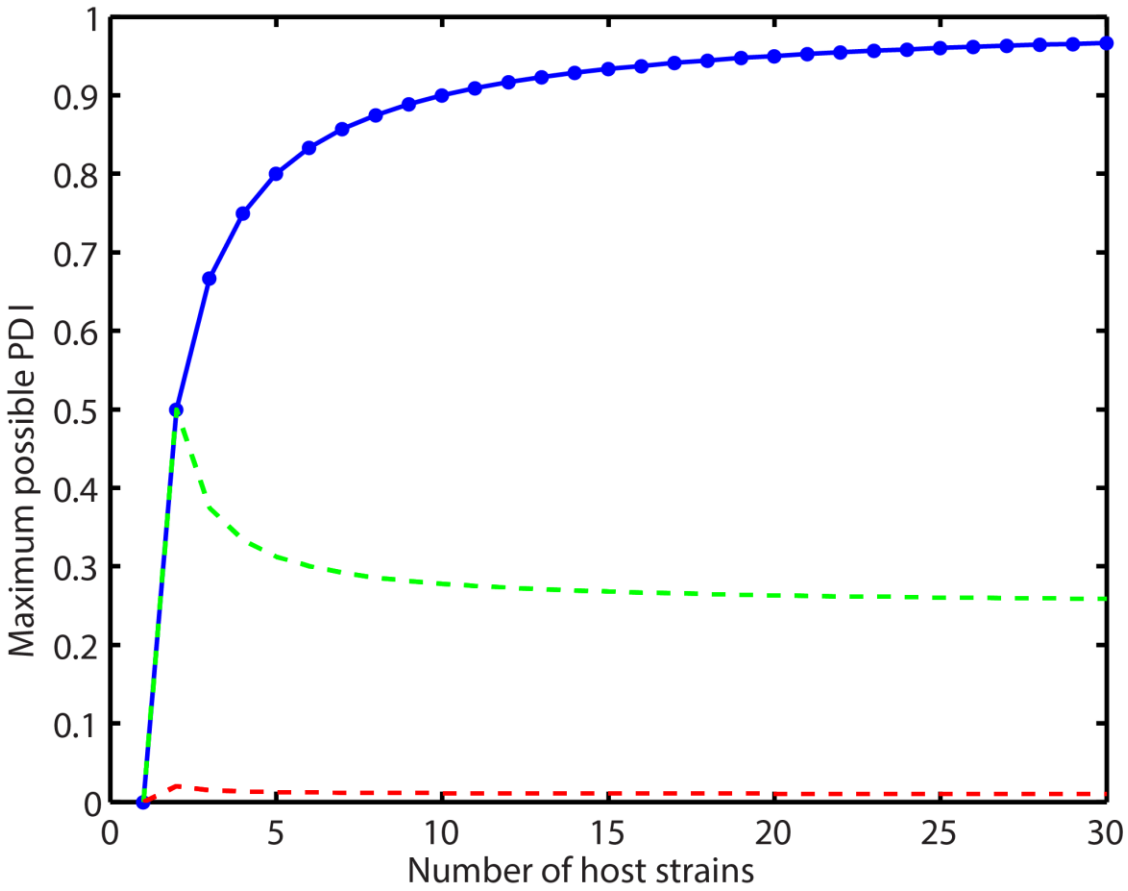


Figure 2.2. Maximum possible PDI changes with the number of host strains. The maximum attainable PDI is determined by the number of host strains, the evenness of the host abundances and requires all host strains are immune to all viral strains. Maximum PDI increases towards one when all hosts have equal abundance (blue). When one host dominates, for example 50% of the population (green) or 90% of the population (red), and all other hosts have equal abundance, the maximum PDI is significantly reduced.

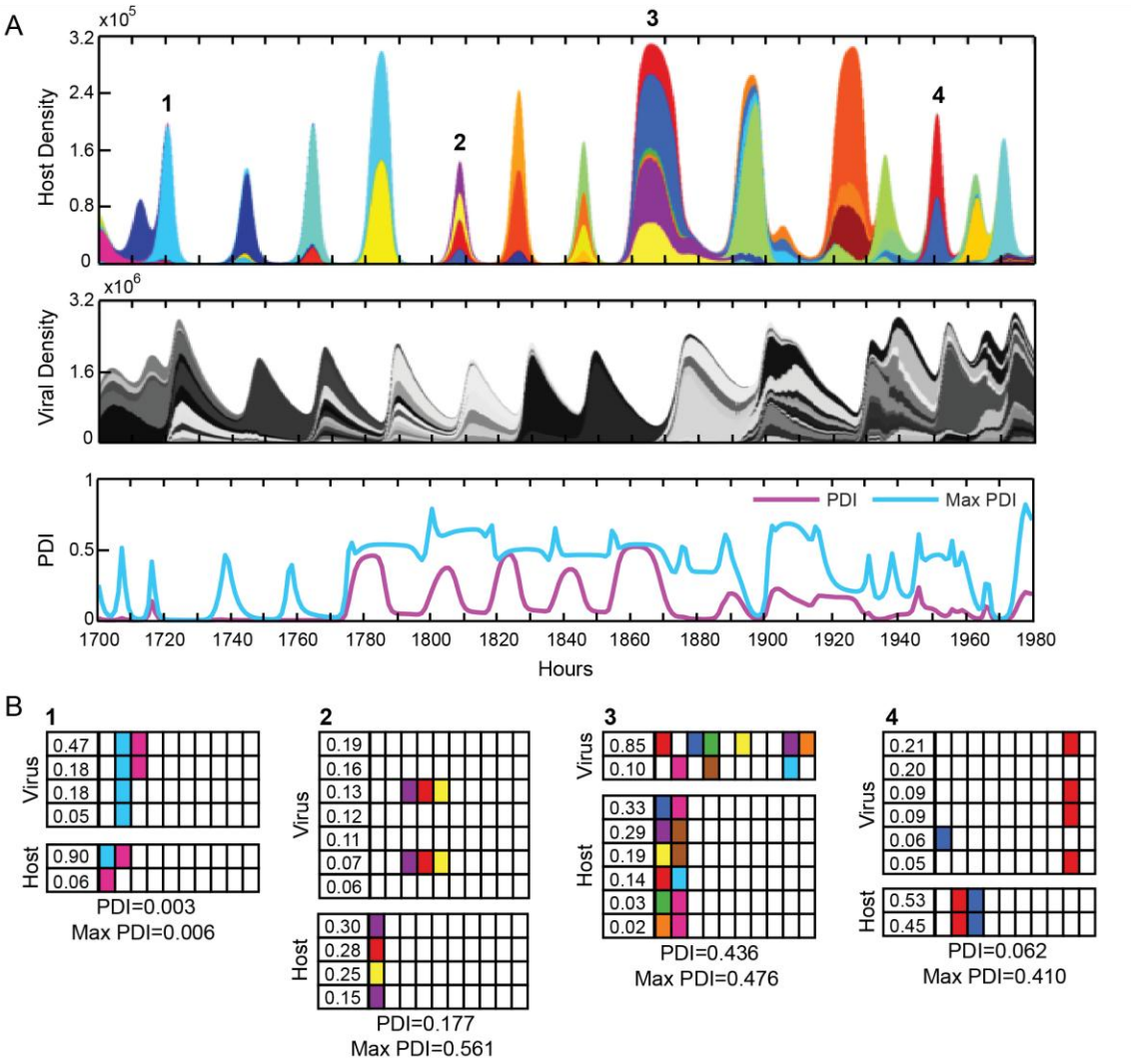


Figure 2.3. **Host populations exhibit periods of different coevolutionary dynamics.** (A) Population dynamics of the host (top) and virus (middle) and PDI (bottom) from a representative simulation. Each color represents a host or viral strain with a unique spacer or protospacer set and is proportional in height to the strain proportion in the population; colors repeat when not touching. (B) Spacer-protospacer matches between major host and viral strains at four time points as examples of single-strain dominance (**1**), coalitions with low immunity (**2**), coalitions with high PDI (**3**), and coalitions with high immunity but low PDI (**4**). The spacer and protospacer composition of each host or viral strain, respectively, is listed horizontally. The number in the first column indicates the proportion of each strain in the population, while the remaining boxes represent the spacer or protospacer state. Host strains making up less than 2% and viral strains making up less than 5% of the population, which only have minor impact on the calculated PDI, are omitted for space. Matching colors in host and viral boxes indicate a spacer-protospacer match. White boxes are spacers or protospacers without a match. Model parameters are standard parameters in Table 2.2.

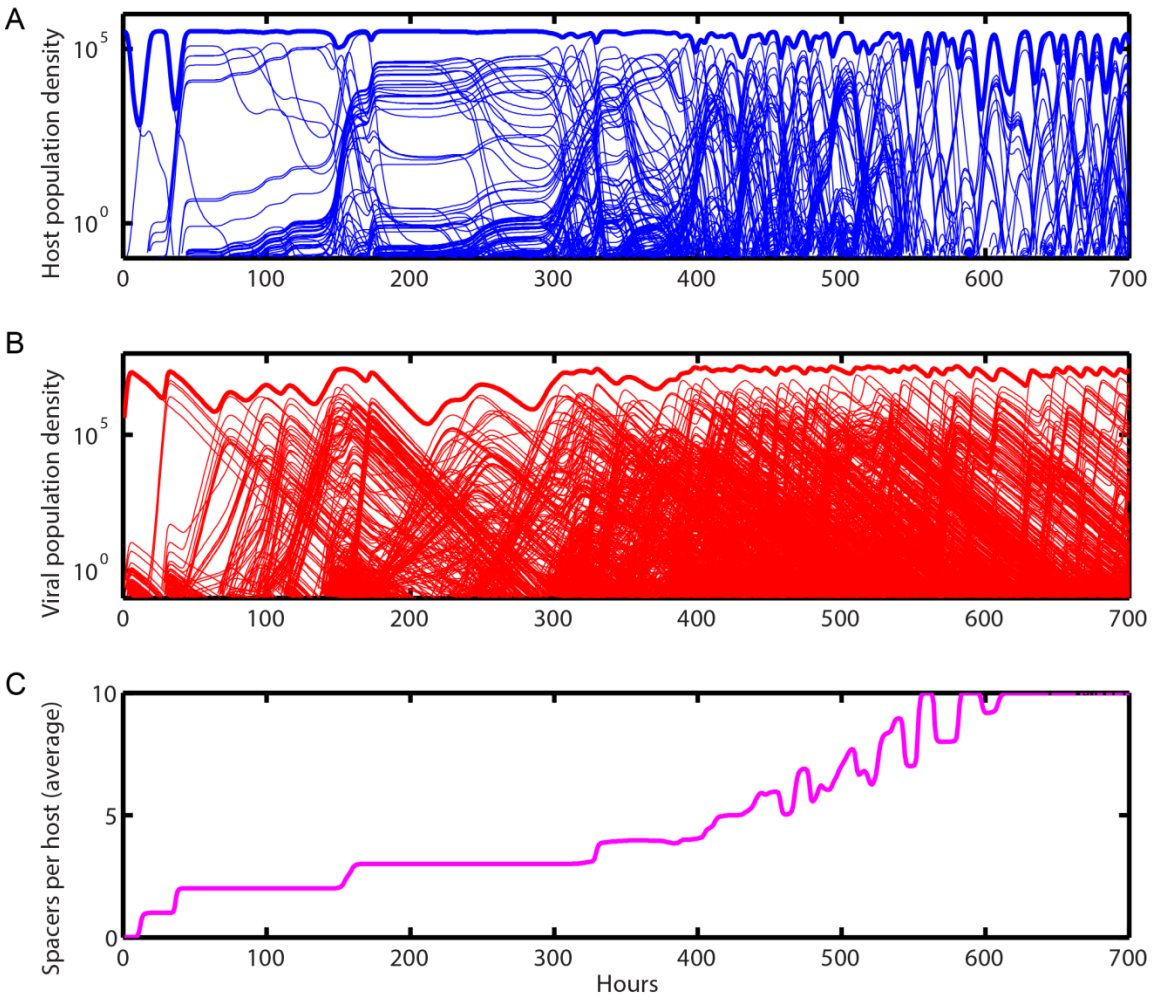


Figure 2.4. Early time course of a representative simulation with standard parameters listed in Table 2.2. Despite seeding with a single host and viral strain, many strains rapidly appear as result of the ever-changing immunity structure. Thick lines at the top of panels A and B are total population density; thin lines are population density of individual host strains (blue lines, A) and viral strains (red lines, B). During the initial hours there is more defined population strain structure when the average spacers per host is low (C).

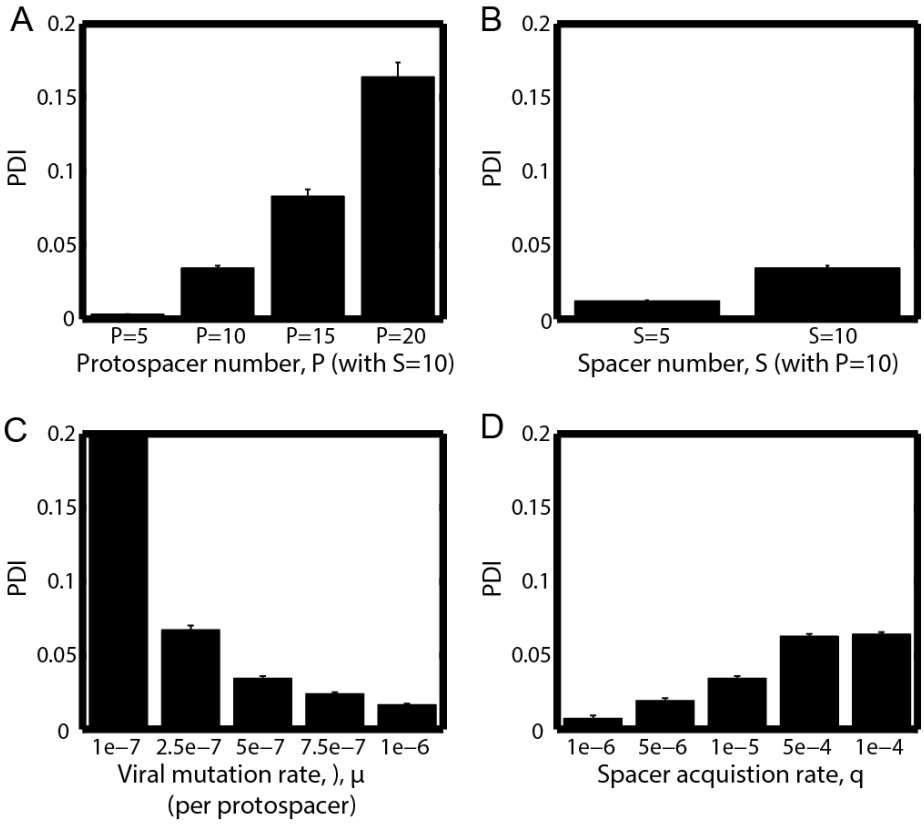


Figure 2.5. PDI is elevated at high protospacer number and low viral mutation rate. Measured PDI in numerical simulations is shown with varying (A) protospacer number; (B) host acquisition rate; (C) viral mutation rate; (D) spacer number. Bars (and lines) represent mean (and SEM) of PDI of replicate simulations, with each replicate represented by the median value across the final 500 hours of that single simulation. Unless varied, parameters are S=10, P=10, q=10^{-5}, μ=5×10^{-7}. Using analysis of variation for unbalanced data all pairwise comparisons of mean PDI are significant at $p<0.001$ except comparisons between: μ=5×10^{-7} and μ=7.5×10^{-7} ($p<0.01$) in (C); μ=7.5×10^{-7} and μ=10^{-6} (not significant) in (C); and q=5×10^{-5} and q=10^{-4} (not significant) in (D).

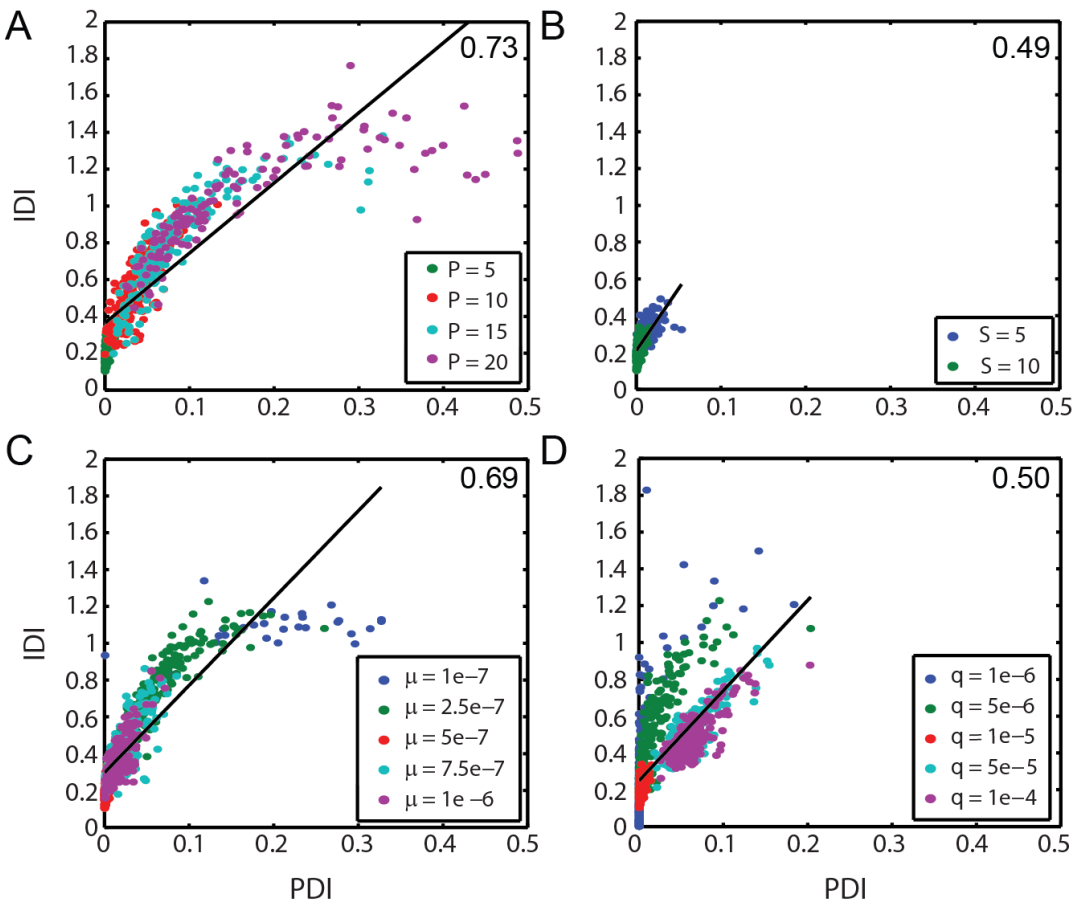


Figure 2.6. PDI is positively correlated with IDI. Each point is the median from last 500 hours of a single simulation varying (A) protospacer number, *P*; (B) spacer number, *S*; (C) viral mutation rate, μ; (D) host spacer acquisition rate, *q*. Unless varied, *S*=10, *P*=10, $q=10^{-5}$, $\mu=5\times10^{-7}$. r^2 correlation coefficients, noted in the upper-right corner of figure panels, of all comparisons are significant at $p<0.001$. Correlations are depicted with solid black lines.

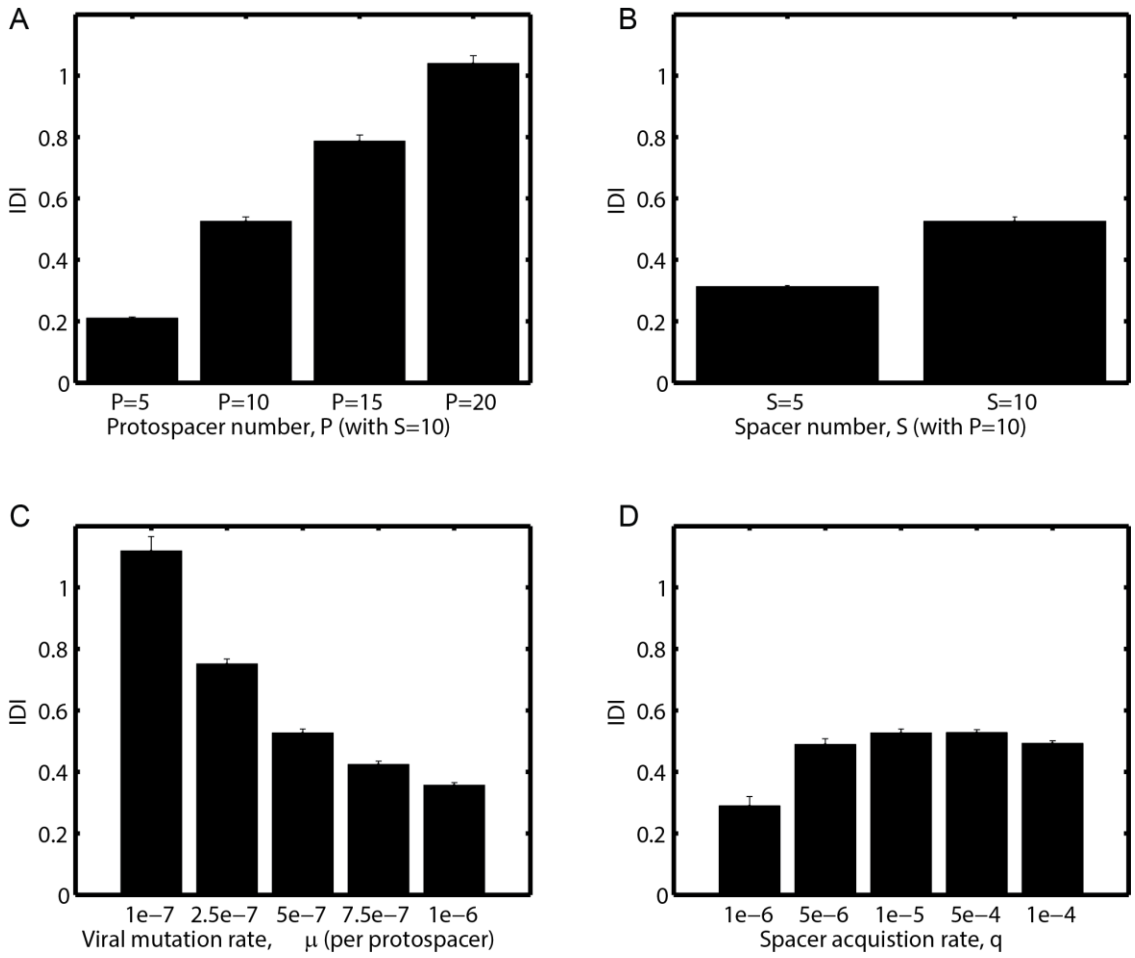


Figure 2.7. IDI varies with: (A) protospacer number; (B) spacer number; (C) viral mutation rate; (D) spacer acquisition rate. Unless varied, S=10, P=10, q=10^{-5}, μ=5x10^{-7}. Bars (and lines) are mean (and SEM) of IDI of replicate simulations, with each replicate represented by the median value across the last 500 hours. Using analysis of variation for unbalanced data all pairwise comparisons of mean PDI are significant at p<0.001 except in (D) where all pairwise comparison with q>1e-6 (not significant).

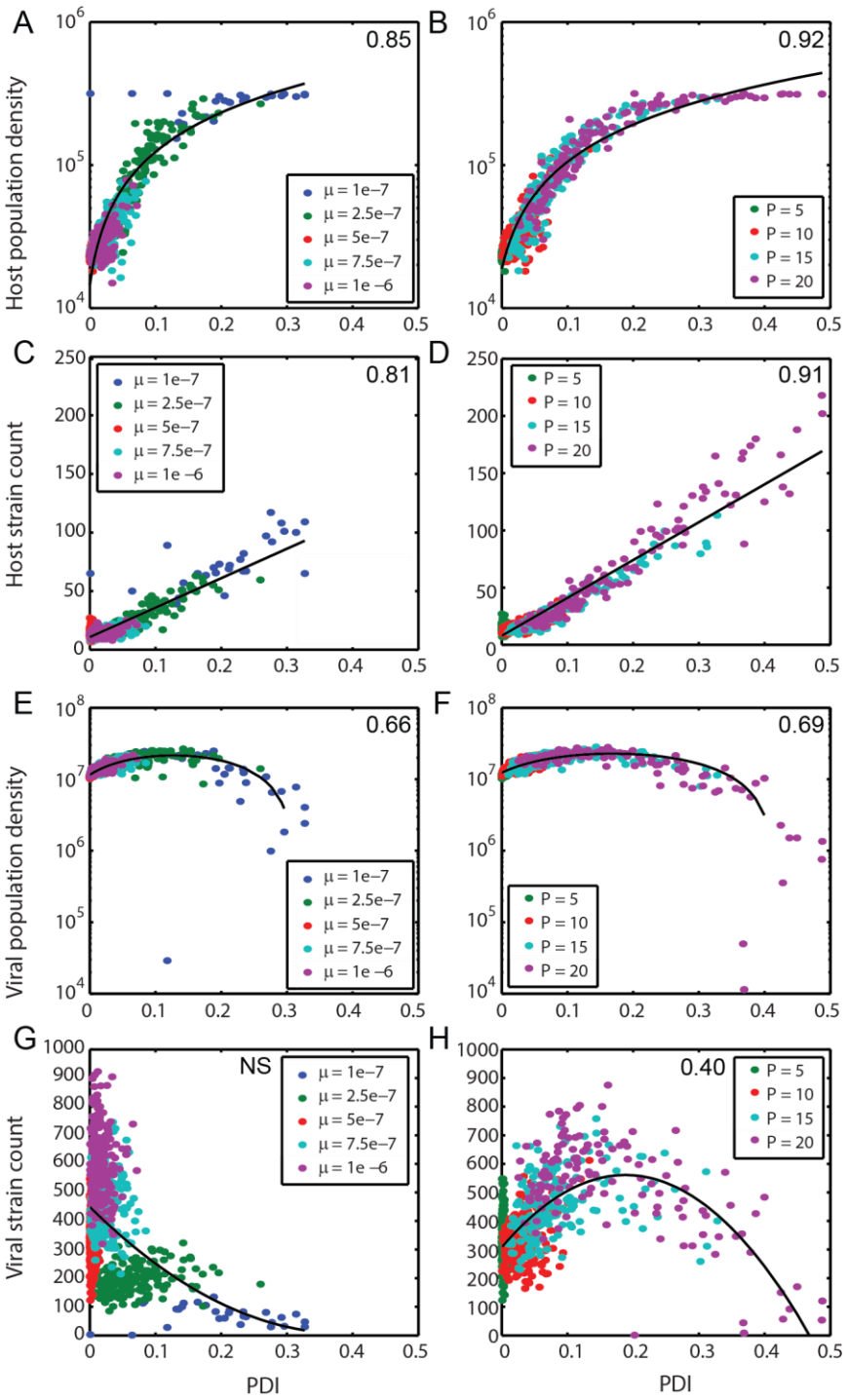


Figure 2.8. PDI and population measures when varying mutation rate and protospacer number. High PDI is associated with (A-B) increases in host population density, (C-D) increases in host strain count, non-monotonic changes (E-F) in viral population density, and non-monotonic changes (G-H) in viral strain count. Left column is varying mutation rate, μ, and right column is varying protospacer number, *P*. Unless varied, parameters are *S*=10, *P*=10, $q=10^{-5}$, $\mu=5\times10^{-7}$. Each point is the median from the last 500 hours of a single simulation. Note that both viral

Figure 2.8. (cont.)

population density and viral strain count are unimodally related to PDI, with the lowest levels of both viral population density and strain count occurring at high PDI. The Spearman rank correlation coefficients of all comparisons, noted in the upper right corner of Figure panels, are significant at $p<0.001$. These relationships, including those that are non-monotonic, are discussed further in the main text.

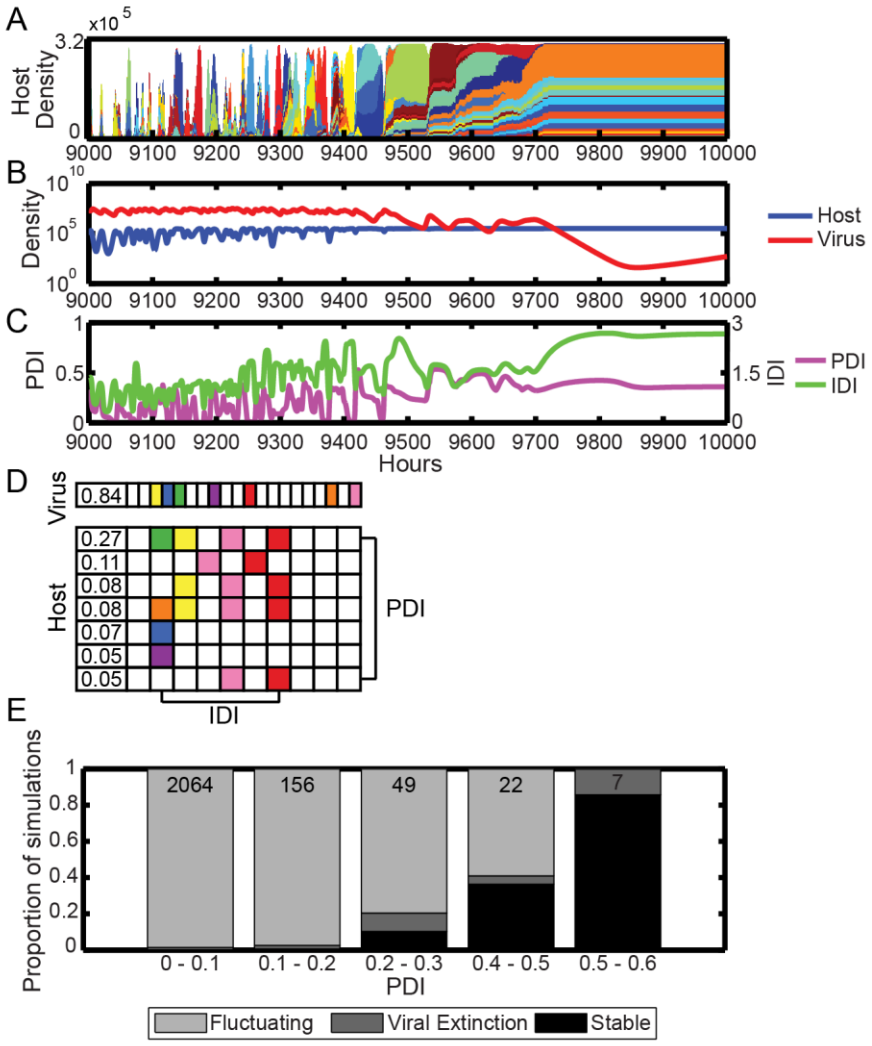


Figure 2.9. Host stability in a high DI population. (A) Plot of host dynamics for a representative model simulation containing an extended period of host population stability. Each color represents a host strain with a unique spacer set and color height is equal to population proportion of the strain; colors repeat when not touching. (B) Total population density in log-scale of host (blue) and virus (red) strains. (C) PDI (magenta, left y-axis) and IDI (green, right y-axis) metrics. (D) Spacer-protospacer matches at 9800 hours. The spacer and protospacer composition of each host or viral strain, respectively, is listed horizontally. The number in the first column indicates the proportion of each strain in the population, while the remaining boxes represent the spacer or protospacer state. Strains making up less than 5% of the population are omitted for clarity. (E) Numbers at the top of each bar designate the total number of simulations in each bin. A simulation is denoted as "stable" when the host population remains above $3x10^5$ (close to carrying capacity) for at least 100 consecutive hours, and as "viral extinction" if the simulation ends prior to the designated endpoint due to reaching a viral population size below our density cutoff of 0.1/mL. Comparisons of subsampled data for stable and viral extinction show significant differences between means of all PDI bins (except between 0-0.1 and 0.1-0.2 stable simulations) and all IDI bins (except between 1.8-2.4 and 2.4-3.0 stable simulations).

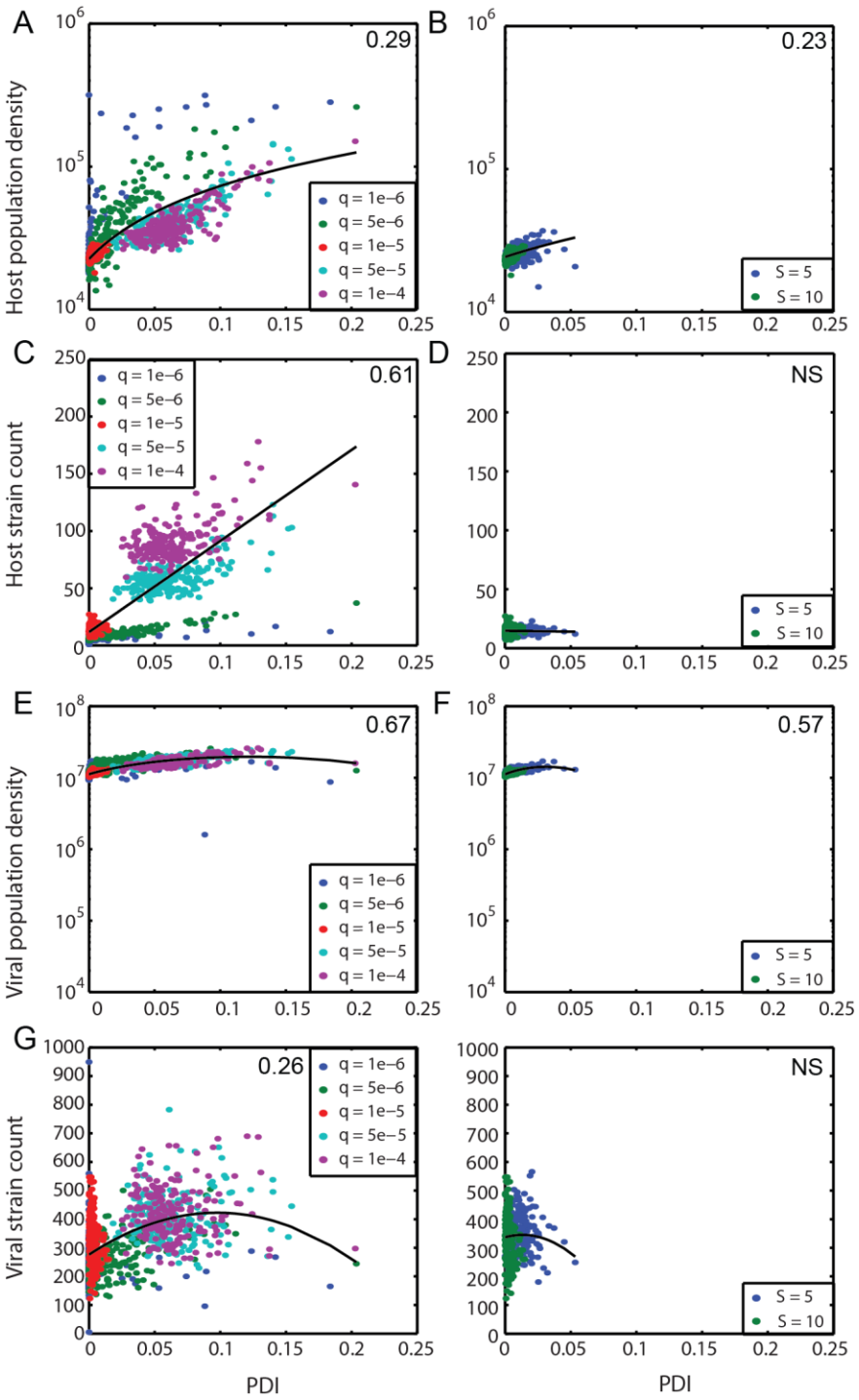


Figure 2.10. PDI and population measures when spacer acquisition rate and spacer number are varied. PDI is only weakly correlated, if at all, with host population density (A-B), host strain count (C-D), viral population density (E-F) and viral strain count (G-H) across variation in spacer acquisition rate, q (left column), and spacer number, S (right column). Unless varied, S=10, P=10, q=10^{-5}, μ=5x10^{-7}. Each point represents the median of the last 500 hours in a single simulation. Linear R^2 correlation coefficients (A-D) and quadratic R^2 correlation coefficients (E-H), noted in

Figure 2.10. (cont.)

the figure panels, of all comparisons are significant at $p<0.001$ except PDI with host strain count (in D) and PDI with viral strain count (in G) when spacer acquisition rate is varied. Correlations are depicted with solid black lines (A-D) and curves (E-H).

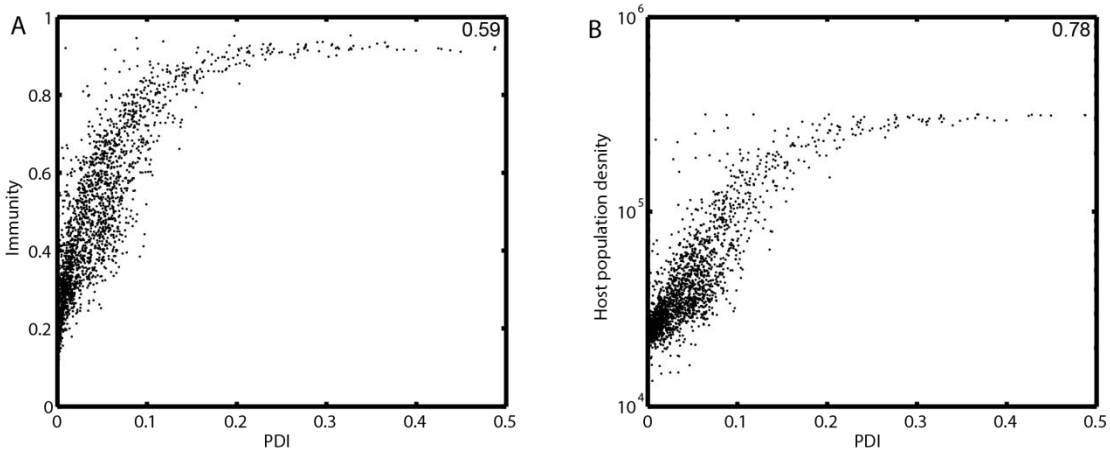


Figure 2.11. Low PDI (< 0.2) is correlated with increases in immunity (A) and host population density (B). At high PDI (> 0.2) immunity (A) and host population density (B) are uniformly high. Each point represents the median of the last 500 hours of a single simulation; all parameter sets from Table 2.1 are included. R^2 correlation coefficients (A) 0.59 and (B) 0.78 are significant at $p<0.001$.

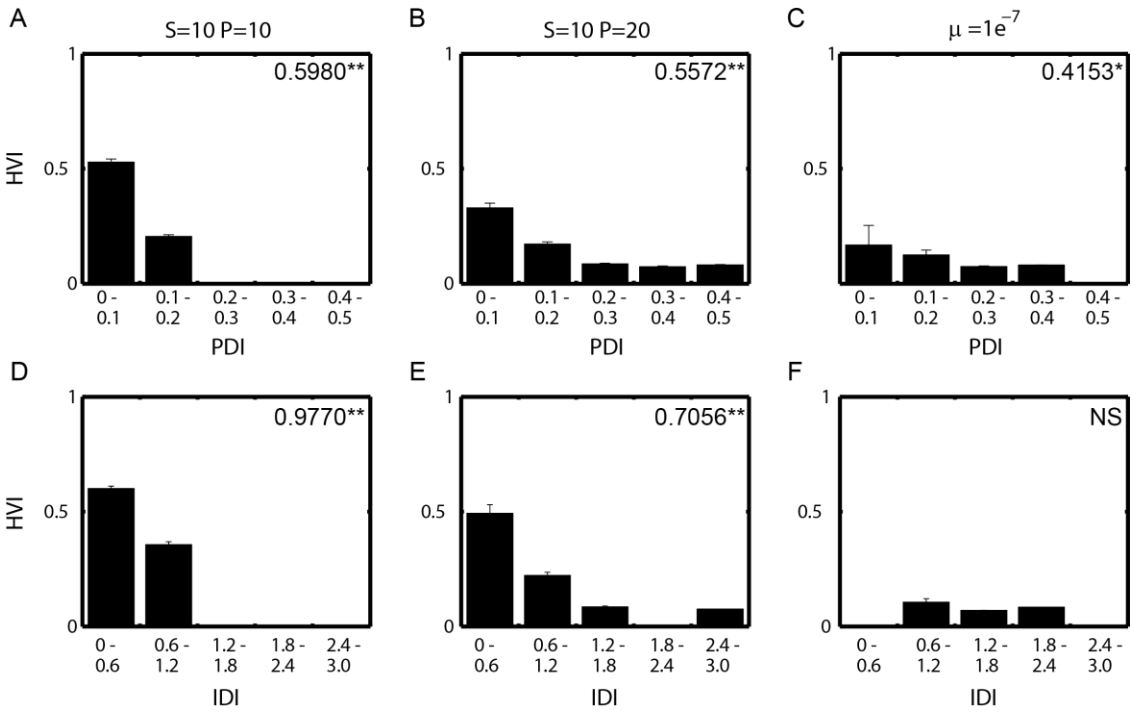


Figure 2.12. **HVI decreases with increasing PDI (A-C) and IDI (D-F).** PDI values binned by 0.1; IDI values binned by 0.6. Bars (and lines) are mean (and SEM) of median HVI across the last 500 hours of each replicate simulation from a pool of 100 simulations per parameter set. Parameters for each panel are (A,D) S=10, P=10, q=10^{-5}, μ=5x10^{-7}; (B,E) S=10, P = 20, q=10^{-5}, μ=5x10^{-7}; (C,F) S=10, P=10, q=10^{-5}, μ = 10^{-7}. All other parameters as listed in Table 2.2. R^2 values (data not binned) are noted in each panel with *, p < 0.01,;**, p << 0.001; NS, not significant.

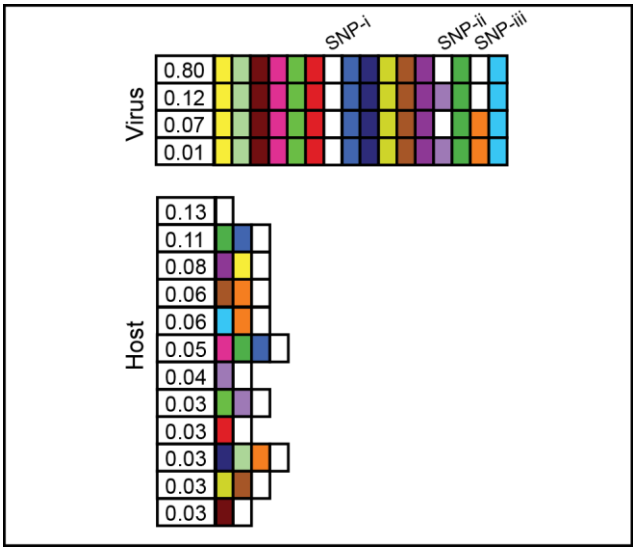


Figure 2.13. PDI and IDI estimated in a population of *Streptococcus thermophilus* and its phage 2972. Data from [53] (accession number SRA049615). Virus protospacer and host spacer states are shown after one week of experimental coevolution. Matching colors in host and viral boxes indicate a spacer-protospacer match. White boxes are spacers or protospacers without a match. All viruses are shown; host strains that make up less than 3% of the population are not shown for clarity. Protospacer positions for which there is no match between any virus and the hosts shown are omitted.

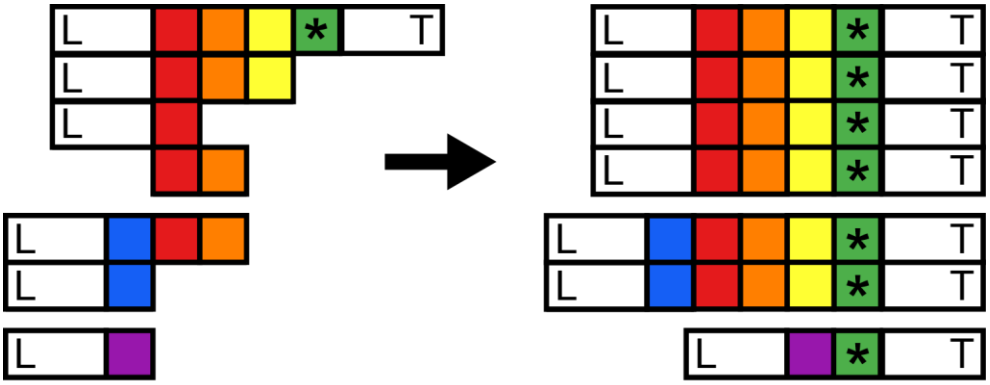


Figure 2.14. **Example of methodology of CRISPR locus reconstruction from sequencing reads.** Each color represents a unique spacer. Each horizontal row on the left shows the spacer content of a single read; its corresponding row on the right shows the inferred complete spacer content. The spacer marked with an asterisk is not present in the ancestral host but has become fixed in the current population. L, leader sequence; T, spacers present in ancestral host.

Table 2.1. Summary of simulated population outcomes. Summary of the population outcomes (complete, viral extinction, unfilled locus) of simulations for each parameter set.

S	*P*	μ	*q*	Complete	Viral Extinction*	Unfilled Locus	Total	Simulation Length (h)
5	10	5.0E-07	1.0E-05	200	0	0	200	2500
10	5	5.0E-07	1.0E-05	200	0	0	200	2500
10	10	5.0E-07	1.0E-05	199	0	1	200	2500
10	15	5.0E-07	1.0E-05	183	2	15	200	10000
10	20	5.0E-07	1.0E-05	132	15	53	200	10000
10	10	1.0E-07	1.0E-05	29	4	167	200	2500
10	10	2.5E-07	1.0E-05	173	0	27	200	2500
10	10	7.5E-07	1.0E-05	200	0	0	200	2500
10	10	1.0E-06	1.0E-05	199	0	1	200	2500
10	10	5.0E-07	1.0E-06	148	16	36	200	2500
10	10	5.0E-07	5.0E-06	198	0	2	200	2500
10	10	5.0E-07	5.0E-05	200	0	0	200	2500
10	10	5.0E-07	1.0E-04	200	0	0	200	2500

* Viral population falling below density cutoff (0.1/mL) during the last 500 hours of the simulation.

Table 2.2. Model parameters. Description of parameters including symbol and value used for simulation of the model.

Parameter	Description	Standard value	Other values
p	CRISPR failure probability	1.0e-05	--
q	spacer acquisition probability	1.0e-05	1.0e-06, 5.0e-06, 5.0e-05, 1.0e-04
r	growth rate (1/h)	1	--
K	carrying capacity (1/mL)	1.0e5.5	--
β	burst size	50	--
Φ	adsorption rate (mL/h)	1.0e-07	--
m	viral decay rate (1/h)	0.1	--
μ	mutation rate	5.0e-07	1.0e-07, 2.5e-07, 7.5e-07, 1.0e-06
ρ	density cutoff (1/mL)	0.1	--
S	number of spacers	10	5
P	number of protospacers	10	5, 15, 20

Table 2.3. Linear - quadratic model comparisons. Summary of the R^2 computation for Figure 2.8E-H and Figure 2.10E-H and choice of model fit using AIC.

Parameter	Type	Description	R^2	AIC	p	Figure
μ	linear	PDI vs. viral population density	0.10	26460	<0.001	2.8E
	quadratic	PDI vs. viral population density	0.66	25674	<0.001	2.8E
P	linear	PDI vs. viral population density	0.01	23960	<0.001	2.8F
	quadratic	PDI vs. viral population density	0.69	23129	<0.001	2.8F
μ	linear	PDI vs. viral strain count	0.19	10425	<0.001	2.8G
	quadratic	PDI vs. viral strain count	0.20	10422	0.02	2.8G
P	linear	PDI vs. viral strain count	0.06	8966	<0.001	2.8H
	quadratic	PDI vs. viral strain count	0.40	8648	<0.001	2.8H
q	linear	PDI vs. viral population density	0.60	30410	<0.001	2.10E
	quadratic	PDI vs. viral population density	0.67	30229	<0.001	2.10E
S	linear	PDI vs. viral population density	0.48	11988	<0.001	2.10F
	quadratic	PDI vs. viral population density	0.57	11913	<0.001	2.10F
q	linear	PDI vs. viral strain count	0.20	11382	<0.001	2.10G
	quadratic	PDI vs. viral strain count	0.26	11313	<0.001	2.10G
S	linear	PDI vs. viral strain count	0.00	4575	0.80	2.10H
	quadratic	PDI vs. viral strain count	0.00	4575	0.17	2.10H

References

[1] Al-Attar, S., Westra, E.R., et al. (2011). Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. *Biol. Chem. 392* (4), 277–289.

[2] Andersson, A.F. and Banfield, J.F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science 320* (5879), 1047–1050.

[3] Angly, F.E., Felts, B., et al. (2006). The marine viromes of four oceanic regions. *PLoS Biol. 4* (11), e368.

[4] Ayala, F.J. and Campbell, C.A. (1974). Frequency-dependent selection. *Annu. Rev. Ecol. Syst. 5* (1), 115–138.

[5] Barrangou, R., Fremaux, C., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science 315* (5819), 1709–1712.

[6] Barrangou, R. and Horvath, P. (2012). CRISPR: new horizons in phage resistance and strain identification. *Annu. Rev. Food Sci. Technol. 3* , 143–162.

[7] Barrick, J.E., Yu, D.S., et al. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature 461* (7268), 1243–1247.

[8] Bikard, D. and Marraffini, L.A. (2012). Innate and adaptive immunity in bacteria: mechanisms of programmed genetic variation to fight bacteriophages. *Curr. Opin. Immunol. 24* (1), 15–20.

[9] Bohannan, B.J.M. and Lenski, R.E. (1997). Effect of resource enrichment on a chemostat community of bacteria and bacteriophage. *Ecology 78* (8), 2303–2315.

[10] Breitbart, M. and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends Microbiol. 13* (6), 278–284.

[11] Brouns, S.J.J., Jore, M.M., et al. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science 321* (5891), 960–964.

[12] Buckling, A. and Rainey, P.B. (2002). Antagonistic coevolution between a bacterium and a bacteriophage. *Proc. R. Soc. B Biol. Sci. 269* (1494), 931–936.

[13] Buckling, A., Wei, Y., et al. (2006). Antagonistic coevolution with parasites increases the cost of host deleterious mutations. *Proc. R. Soc. B Biol. Sci. 273* (1582), 45–49.

[14] Childs, L.M., Held, N.L., et al. (2012). Multiscale model of CRISPR-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin. *Evolution 66* (7), 2015–2029.

[15] Chopin, M.-C., Chopin, A., et al. (2005). Phage abortive infection in lactococci: variations on a theme. *Curr. Opin. Microbiol. 8* (4), 473–479.

[16] Cui, Y., Li, Y., et al. (2008). Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. *PLoS One 3* (7), e2652.

[17] Desai, M.M., Fisher, D.S., et al. (2007). The speed of evolution and maintenance of variation in asexual populations. *Curr. Biol. 17* (5), 385–394.

[18] Deveau, H., Barrangou, R., et al. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol. 190* (4), 1390–1400.

[19] Deveau, H., Garneau, J.E., et al. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol. 64* , 475–493.

[20] Drake, J.W. (2009). Avoiding dangerous missense: thermophiles display especially low mutation rates. *PLoS Genet 5* (6), e1000520.

[21] Fisher, R.A. (1930). *The Genetical Theory Of Natural Selection*. Oxford University Press, Oxford.

[22] Forde, S.E., Thompson, J.N., et al. (2008). Coevolution drives temporal changes in fitness and diversity across environments in a bacteria-bacteriophage interaction. *Evolution 62* (8), 1830–1839.

[23] Gandon, S. and Vale, P.F. (2014). The evolution of resistance against good and bad infections. *J. Evol. Biol. 27* (2), 303–312.

[24] Gerrish, P.J. and Lenski, R.E. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica 102–103* (1–6), 127–144.

[25] Grissa, I., Vergnaud, G., et al. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res. 35* (Suppl 2), W52–W57.

[26] Haerter, J.O. and Sneppen, K. (2012). Spatial structure and Lamarckian adaptation explain extreme genetic diversity at CRISPR locus. *mBio 3* (4), e00126-12.

[27] He, J. and Deem, M.W. (2010). Heterogeneous diversity of spacers within CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats). *Phys. Rev. Lett. 105* (12), 128102.

[28] Held, N.L., Childs, L.M., et al. (2013). CRISPR-Cas systems to probe ecological diversity and host–viral interactions. In R. Barrangou and J. van der Oost, eds., *CRISPR-Cas Systems*. Springer Berlin Heidelberg, 221–250.

[29] Held, N.L., Herrera, A., et al. (2010). CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS One 5* (9), e12988.

[30] Held, N.L., Herrera, A., et al. (2013). Reassortment of CRISPR repeat-spacer loci in *Sulfolobus islandicus*. *Environ. Microbiol. 15* (11), 3065–3076.

[31] Horvath, P. and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. *Science 327* (5962), 167–170.

[32] Horvath, P., Romero, D.A., et al. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol. 190* (4), 1401–1412.

[33] Hyman, P. and Abedon, S.T. (2010). Bacteriophage host range and bacterial resistance. *Adv. Appl. Microbiol. 70* , 217–248.

[34] Iranzo, J., Lobkovsky, A.E., et al. (2013). Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. *J. Bacteriol. 195* (17), 3834–3844.

[35] Jiang, W., Maniv, I., et al. (2013). Dealing with the evolutionary downside of CRISPR immunity: Bacteria and beneficial plasmids. *PLoS Genet. 9* (9), e1003844.

[36] Labrie, S.J., Samson, J.E., et al. (2010). Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol. 8* (5), 317–327.

[37] Lenski, R.E. (1988). Experimental studies of pleiotropy and epistasis in *Escherichia coli*. I. Variation in competitive fitness among mutants resistant to virus T4. *Evolution 42* (3), 425–432.

[38] Lenski, R.E. and Levin, B.R. (1985). Constraints on the coevolution of bacteria and virulent phage: a model, some experiments, and predictions for natural communities. *Am. Nat. 125* (4), 585–602.

[39] Levin, B.R. (2010). Nasty viruses, costly plasmids, population dynamics, and the conditions for establishing and maintaining CRISPR-mediated adaptive immunity in bacteria. *PLoS Genet 6* (10), e1001171.

[40] Levin, B.R., Antonovics, J., et al. (1988). Frequency-dependent selection in bacterial populations [and discussion]. *Philos. Trans. R. Soc. Lond. B Biol. Sci. 319* (1196), 459–472.

[41] Levin, B.R., Moineau, S., et al. (2013). The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. *PLoS Genet. 9* (3), e1003312.

[42] Lopez-Pascua, L. d C. and Buckling, A. (2008). Increasing productivity accelerates host-parasite coevolution. *J. Evol. Biol. 21* (3), 853–860.

[43] Makarova, K.S., Aravind, L., et al. (2011). Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol. Direct 6* , 38.

[44] Marraffini, L.A. and Sontheimer, E.J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet 11* (3), 181–190.

[45] Mojica, F.J.M., Díez-Villaseñor, C., et al. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol. 60* (2), 174–182.

[46] Mojica, F.J.M., Díez-Villaseñor, C., et al. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology 155* (3), 733–740.

[47] Molineux, I.J. (1991). Host-parasite interactions: recent developments in the genetics of abortive phage infections. *New Biol. 3* (3), 230–236.

[48] van der Oost, J., Jore, M.M., et al. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci. 34* (8), 401–407.

[49] Paez-Espino, D., Morovic, W., et al. (2013). Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat. Commun. 4* , 1430.

[50] Rainey, Buckling, et al. (2000). The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends Ecol. Evol. 15* (6), 243–247.

[51] Rho, M., Wu, Y.-W., et al. (2012). Diverse CRISPRs evolving in human microbiomes. *PLoS Genet. 8* (6), e1002441.

[52] Semenova, E., Jore, M.M., et al. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci. 108* (25), 10098–10103.

[53] Sun, C.L., Barrangou, R., et al. (2013). Phage mutations in response to CRISPR diversification in a bacterial population. *Environ. Microbiol. 15* (2), 463–470.

[54] Suttle, C.A. (2007). Marine viruses — major players in the global ecosystem. *Nat. Rev. Microbiol. 5* (10), 801–812.

[55] Terns, M.P. and Terns, R.M. (2011). CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol. 14* (3), 321–327.

[56] Tock, M.R. and Dryden, D.T. (2005). The biology of restriction and anti-restriction. *Curr. Opin. Microbiol. 8* (4), 466–472.

[57] Touchon, M. and Rocha, E.P.C. (2010). The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS One 5* (6), e11126.

[58] Tyson, G.W. and Banfield, J.F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol. 10* (1), 200–207.

[59] Weinberger, A.D., Sun, C.L., et al. (2012). Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput Biol 8* (4), e1002475.

[60] Weinberger, A.D., Wolf, Y.I., et al. (2012). Viral diversity threshold for adaptive immunity in prokaryotes. *mBio 3* (6), e00456-12.

[61] Weitz, J.S., Hartman, H., et al. (2005). Coevolutionary arms races between bacteria and bacteriophage. *Proc. Natl. Acad. Sci. U. S. A. 102* (27), 9535–9540.

[62] Westra, E.R., Swarts, D.C., et al. (2012). The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet. 46* , 311–339.

[63] Wilson, G.G. and Murray, N.E. (1991). Restriction and modification systems. *Annu. Rev. Genet. 25* (1), 585–627.

[64] Winter, C., Bouvier, T., et al. (2010). Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "Killing the Winner" hypothesis revisited. *Microbiol. Mol. Biol. Rev. 74* (1), 42–57.

CHAPTER 3

CRISPR Diversity in the Vaginal Microbiomes of Pregnant Women

Abstract

The vaginal microbiome has a significant impact on womens' health, including connections to preterm birth risk; however, the role of bacteriophages in these communities is largely unstudied. The CRISPR system, a microbial immune system which stores pieces of foreign DNA, has the potential to offer insight into how these microbes interact with their viruses. We employed a network-based method to identify CRISPRs from all possible species in metagenomic data from the vaginal microbiota of ten women over the course of their pregnancies; five of whom were at high risk for preterm delivery and five of whom were low-risk. We identified over 20 types of CRISPRs, with variation in spacer content among individuals. We found evidence of multiple coexisting CRISPR alleles, CRISPR change via spacer addition over time, and CRISPR-linked shifts in protospacer abundance in CRISPRs from *Lactobacillus iners* in one of our subjects, demonstrating that our method is capable of detecting CRISPR variation in metagenomic samples and that CRISPRs are actively interacting with foreign elements in the vaginal microbiome.

Introduction

The human vaginal microbiome consists of a community of microbes with implications for health and disease. Microbial composition has been linked to a range of medical complications such as bacterial vaginosis [21,29], sexually transmitted diseases [17], and risk of preterm birth [14,19,33]. Though studies have not supported a single core vaginal microbiome [22], a series of community groups (called Types I-V) dominated by a single *Lactobacillus* species or a consortium of strict anaerobes have been identified [22]. *Lactobacillus* is traditionally associated with a healthy state, producing lactic acid to maintain appropriate vaginal pH, while a community dominated by anaerobes is diagnostic of bacterial vaginosis [20]. The microbial community can change over time, both short-term fluctuations and in response to major events such as pregnancy. During pregnancy, the vaginal community is characterized by lower species diversity, higher

Lactobacillus abundance, and lower abundance of anaerobic microbes present in Type IV communities [1,26,32]. The vaginal microbial community is also linked to preterm birth; reproductive tract infections are a major cause of preterm birth [10], and a correlation has been found between preterm birth and lower vaginal community diversity [14].

Compared to what is understood about the effects of the bacterial communities of the vagina, relatively little is known about the role of bacteriophages in these communities. Prophage induction in lactobacilli has been posited to contribute to development of bacterial vaginosis by killing off resident lactobacilli, allowing vaginosis-linked species to proliferate [30], and these prophages have also been proposed to be influential agents of horizontal gene transfer in the vaginal microbiome [4]. However, the ecology of phage populations and how bacterial hosts modulate their interactions with these phages through resistance and immunity mechanisms has not been thoroughly studied.

One immunity mechanism which holds promise for investigating bacteria-phage interactions is the CRISPR system, an adaptive microbial immune system which is present in 50% of sequenced bacteria [16]. The CRISPR system consists of arrays of palindromic repeats interspersed with spacers, or short DNA fragments which frequently match segments of foreign genetic elements such as viruses, plasmids, and transposons [18]. The spacer-matched sequences in these elements are referred to as protospacers. Hosts harboring CRISPRs can add new spacers to their repeat-spacer arrays in a polar manner, with the end to which new spacers are added called the leader [3]. The complete repeat-spacer array is transcribed and processed into crRNAs, each of which holds the sequence of an individual spacer. These crRNAs are used to guide protein complexes to corresponding protospacers when the matched element is encountered again. These complexes inactivate and/or degrade the targeted DNA or RNA, preventing infection or transfer of the element [5]. Because they integrate new spacers in the order in which they encounter various foreign elements, CRISPR arrays can be used to trace a chronological history of phage encounters in addition to recording the present immunity of each strain.

To investigate CRISPR presence, diversity, and change over time in the vaginal microbiome, we used a network-based method to extract CRISPR repeat-spacer arrays from vaginal microbiome sequences of ten pregnant women during and after their

pregnancies. We examined CRISPR diversity in each sample, as well as variation in CRISPRs between individuals and over the duration of pregnancy within each subject. We were able to identify CRISPRs in many microbial species and assess within-species CRISPR variation, enabling us to examine spacer variation within and between individuals.

Results

CRISPR presence is widespread in the vaginal microbiome

At least one CRISPR repeat type was detected in 43 out of 53 samples analyzed; only one of the 10 subjects (201) lacked detectable CRISPR loci at all sampled time points (Figure 3.1). Interestingly, all CRISPR-less samples were obtained from subjects with a history of preterm birth. CRISPR-containing reads were typically associated with the most abundant organisms in the sample, most frequently *Lactobacillus iners*, *Lactobacillus crispatus*, or *Gardnerella vaginalis*. However, in samples from subjects 125 and 202, CRISPR reads were most frequently from *Lactobacillus jensenii* and *G. vaginalis*, while the samples were dominated by *L. iners* and *Lactobacillus gasseri*, respectively (Figure 3.1).

Spacer variation among individuals

CRISPR repeat types were further subclassified into subtypes based on spacer content. Multiple instances of individuals harboring the same CRISPR repeat type with different spacer types were identified (Figure 3.2). These occurrences were found in CRISPR loci from *L. crispatus*, *G. vaginalis*, and *Anaerococcus lactolyticus*. In the case of *L. crispatus*, two CRISPR alleles emerged, each consisting of a single shared spacer and three unique spacers. Allele 1A was present in two subjects, while 1B was found in three. Seven unique spacers exist in the *G.vaginalis* type 3B locus, which appears in a single sample, while its other 16 spacers are present in Type 3A, present only in a different subject. The final case, *A. lactolyticus*, has three completely different sets of spacers, with none shared between them; each shows up in a single sample from a different subject. In all cases, no subject had multiple alleles simultaneously, nor did the allele present change over time (Figure 3.2).

Spacer variation over time within an individual

Persistence of a repeat type over multiple time points in an individual was common; out of the nine subjects with CRISPR arrays identified, seven retained at least one repeat type for at least 3 consecutive samples (Figure 3.2). Variation in the spacer content of these arrays, however, was less frequent. Only one subject, 110, exhibited coexistence of multiple CRISPR alleles of the same repeat type. Samples 110_1 and 110_2 contained three distinct *L. iners* CRISPR alleles, each distinguished by an additional leader-end spacer (Figure 3.3). The same subject was also the sole example of change in spacer content over time. After a period of undetectable *L. iners* CRISPRs in late pregnancy and labor, at 6 weeks postpartum, *L. iners* CRISPRs were again detectable, with a single CRISPR allele sharing many spacers of the most common allele present in early pregnancy, but with a novel leader-end spacer (Figure 3.3).

Relationship between CRISPR variation and extrachromosomal elements

To investigate whether changes in CRISPR type, abundance, or spacer content were correlated with changes in matched protospacers, we identified protospacer sequences matching all identified spacers in each sample (Figure 3.4). In sample 110_1, nine protospacers matched by *L. iners* spacers were present at high abundance; at subsequent time points, the abundance of the protospacers was substantially decreased or eliminated (Figure 3.4). However, in other instances where both spacer and matched protospacer are present in the same subject, no apparent correlation exists between abundance of spacer and protospacer (Figure 3.4).

Discussion

Through bioinformatic analysis of a time series of samples from a cohort of pregnant women, we were able to identify CRISPR repeat-spacer arrays from a variety of organisms present in the vaginal microbiome. While CRISPRs exist in a number of different species, with variation in spacer content between strains present in different individuals, we found remarkably limited diversity in spacer content within individuals, with limited evidence of coexistence of multiple CRISPR alleles or change in alleles over time. We also found little correlation between spacer and matched protospacer presence

in the same sample; however, the primary examples of both within-patient CRISPR diversity and change in matched protospacer presence come from the same species in the same subject: *L. iners* from subject 110.

CRISPRs are frequently among the most diverse parts of the genome, with extensive variation of CRISPR alleles found in metagenomes from acid mine drainage biofilms [31], *Yersinia pestis* isolates from Asian plague foci [9,25], a collection of *Streptococcus thermophilus* strains [13], isolates from a single hot spring population of *Sulfolobus islandicus* [11,12], and the oral and gut microbiomes [23]. Modeling of CRISPR-mediated host-virus interactions has also predicted the emergence and maintenance of CRISPR diversity [6,7]. Given this history of diversity, we anticipated finding variation within individual subjects, despite the previously observed limited species-level diversity of the vaginal microbiome, especially during pregnancy [1,26,32]. However, we only found multiple CRISPR alleles coexisting and changing over time in a single subject. This may indicate that these communities have low exposure to phages and other extrachromosomal elements, or that they interact with a limited pool of elements, such that most have already acquired spacers against most of the common elements. The upheaval of the vaginal community associated with pregnancy may also significantly disturb the viral community, leading to an environment with limited phage pressure. This limited exposure would lower selection for strains which have added novel spacers, preventing them from increasing in frequency and becoming detectable in our metagenomic samples. It is also possible that some of the many other mechanisms of resisting viral infection [15] predominate in these communities.

Despite low diversity within individuals, heterogeneity between subjects was more common; of the three repeat types present in more than one individual, two had a different spacer set in each person; the third had two distinct alleles, with one or the other present in each subject. This is consistent with broader examination of variation in the human microbiome, which finds highly personalized communities among individuals at multiple body sites [8]. However, despite this heterogeneity, in most cases where multiple subjects contain the same dominant species, no variation in CRISPR content is observed. As with the lack of within-subject variation, this low diversity points to an

environment where either a limited pool of phage is encountered by these bacteria, or CRISPR immunity is not the preferred mechanism of evading viral infection.

A previous examination of CRISPRs in 61 vaginal microbiomes of non-pregnant women [23] identified 4 known CRISPR repeat types, as well as a novel repeat. The 21 different CRISPR repeats found by our analysis include three of these, all from *Lactobacillus* species, and a repeat type similar to Rho et al.'s *A. lactolyticus* repeat. We did not identify the novel repeat type in our dataset. Rho et al. identified the three *Lactobacillus* repeats in roughly equal numbers of metagenomes (28, 31, and 33 out of 61); by contrast, we found that while two of the repeat types were found in 17 out of 53 samples, always co-occurring and from five different subjects, the other was found in only four samples, all from one subject. These similarities and differences reinforce that while the vagina is generally colonized by a handful of common species, there can still be notable variance in community structure, especially when these communities undergo shifts as a result of significant events such as pregnancy.

In addition to low levels of CRISPR variation, we also saw limited correlation between presence of CRISPR-matched elements and their cognate spacers. Elements matched by spacers were found in every sample (Figure 3.4), indicating that these communities do encounter elements to which they have immunity; however, their abundance had little to do with presence or absence of matching spacers in the sample. One notable change in matched element abundance was observed in subject 110. Sequences matching nine spacers contained in *L. iners* CRISPR arrays were present at relatively high abundance in the first sample obtained; however, at later time points the abundance of these sequences drops sharply, in some cases becoming undetectable at this sequencing depth. The even abundance of each of these protospacers implies that they may all be present in a single element, though they cannot be conclusively linked; as such, we may have captured the recent introduction or increase in abundance of a phage, plasmid or other element into this community. Notably, a novel spacer was added to the *L. iners* CRISPR array in the final sample from this subject, taken six weeks postpartum. This spacer appears at high abundance in the first sample and then decreases, undergoing identical dynamics to the nine previously described and implying this spacer may match the same element. This is

consistent with an active *L. iners* CRISPR system adding a new spacer to a recently encountered element. While the variation we detected in this dataset was limited, this observation shows that our method of identifying CRISPR array variants in metagenomic data using network analysis is sensitive enough to pick up CRISPR variants through time.

Methods

Sample collection

Subjects were recruited from the patient population at the Mayo Clinic, Rochester, MN. Five subjects with a history of preterm birth (defined as at least one pregnancy resulting in preterm delivery) and five subjects with no history of preterm birth (defined as at least one successful pregnancy, and no pregnancies ending in preterm delivery) were selected. All subjects were Caucasian and had resided in Rochester for a minimum of 23 years. Consent to sample collection was obtained from each patient at each collection time. Swabs were collected from the posterior fornix of each subject at up to six time points: gestational week 8-12, 17-21, 26-30, 35-38, labor, and six weeks postpartum (Table 3.1).

Sequencing, processing and taxonomic assignment

DNA was extracted from each sample, and Illumina HiSeq DNA sequencing was performed, producing 100 bp paired end reads. Quality filtered sequence data cleaned of human DNA contamination was kindly provided by Fang Yang and Bryan White of the Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign. Sequences shorter than 50 bp or with an average quality score below 20 were discarded. Reads were assigned to species via blastn (e=1e-5) against the Human Microbiome Project urogenital reference database (downloaded Nov. 15, 2013).

CRISPR identification via network analysis

Potential CRISPR repeat sequences in each sample were identified using CRASS [27]. Potential repeats were compared via blastn [2] to the NCBI nt database; any sequences which primarily hit non-microbial sequences were classified as false positives and excluded from further analysis. Reads matching remaining potential CRISPR repeats or a list of known *Lactobacillus* CRISPR repeats were identified using blastn (blastn-short,

e<=0.1, >=90% nucleotide identity) . To establish spacer order and detect even rare coexisting CRISPR alleles, spacers were linked as follows: 12-nucleotide “chunks” of DNA flanking each repeat were identified using fuzznuc [24]; up to 8 mismatches to the repeat sequence were allowed to capture degenerate repeats. Due to the palindromic nature of the repeats, this occasionally generated situations where the repeat was matched in both orientations; in these cases, the match with fewer mismatches was kept and the secondary match discarded. Chunks that were a perfect match to the repeat sequence (i.e., from two adjacent repeats or partial repeats) were also discarded. Finally, singleton chunks that perfectly overlap non-singleton chunks by at least 8 bp were removed, to account for rare chunks generated by sequencing error.

Custom bash scripts were used to identify occurrences of two or more chunks on the same read. These were recorded as links, which represent either two ends of the same spacer, or opposite ends of two spacers linked across a repeat. The first type of link was used to identify spacer sequences; the second, to order spacers and identify branch points, i.e. where multiple alleles with partially shared spacer content exist. Linkage networks were analyzed using Cytoscape [28]. Based on average repeat and spacer lengths of species known to inhabit the vagina, links spanning 60 bp or less were considered short links, spanning only a single spacer or pair of adjacent spacers; longer links were considered to span multiple spacers and were not counted when determining coverage of links across a spacer or between adjacent spacers.

All sequences flanked by repeats on both sides were considered potential spacers and were extracted. All spacers identified using the same repeat were considered a repeat type, denoted by a number; variant alleles within a repeat type were designated by adding a letter to their group number (i.e. 2A, 2B, etc.).

Identification of matched protospacers

To determine presence and abundance of protospacer-containing elements matched by our spacer library, we used blastn (blastn_short e<=0.1, >=90% nucleotide identity) to compare all identified spacers from all samples to a database of all metagenome reads. Reads previously used to build CRISPR arrays were excluded from analysis.

Figures and Tables

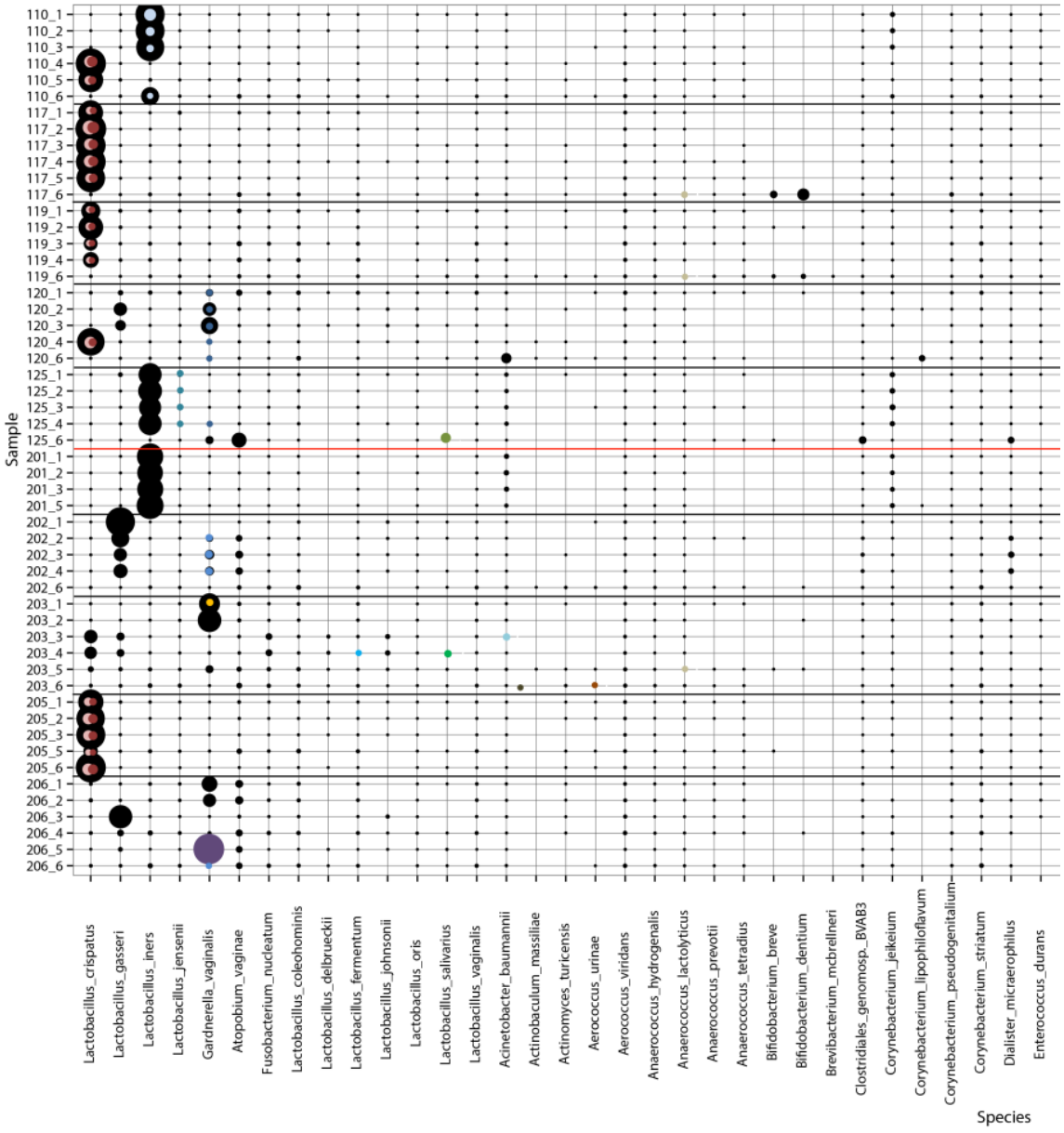
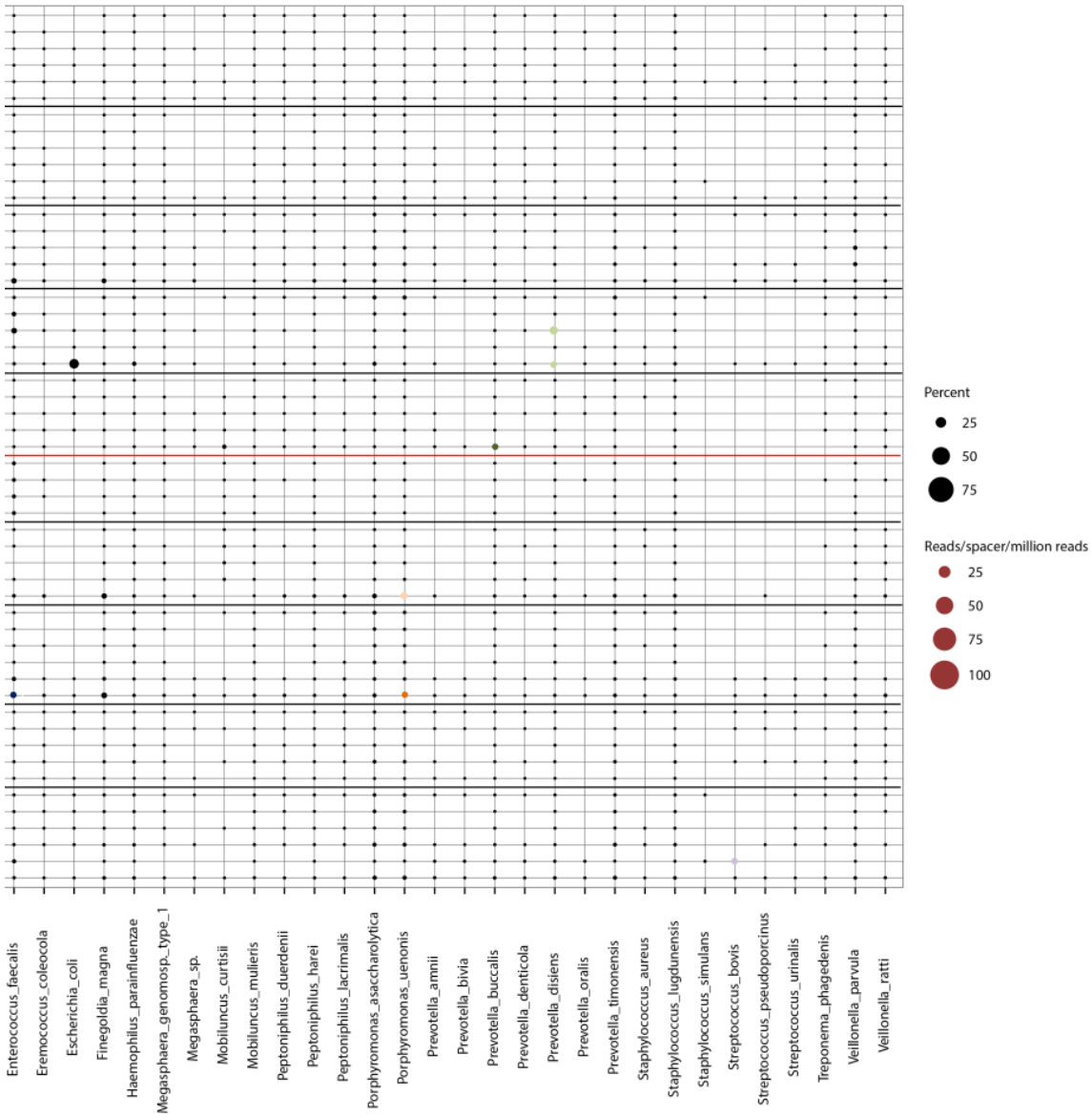


Figure 3.1. Species abundance and CRISPR presence in the vaginal microbiome. Black circles represent the proportion of reads from each sample which correspond to the species; colored circles represent the number of reads matching CRISPR loci per spacer per million reads. Black lines separate individual subjects, while the red line separates low-risk subjects (above) from high-risk subjects (below).

Figure 3.1. (cont.)



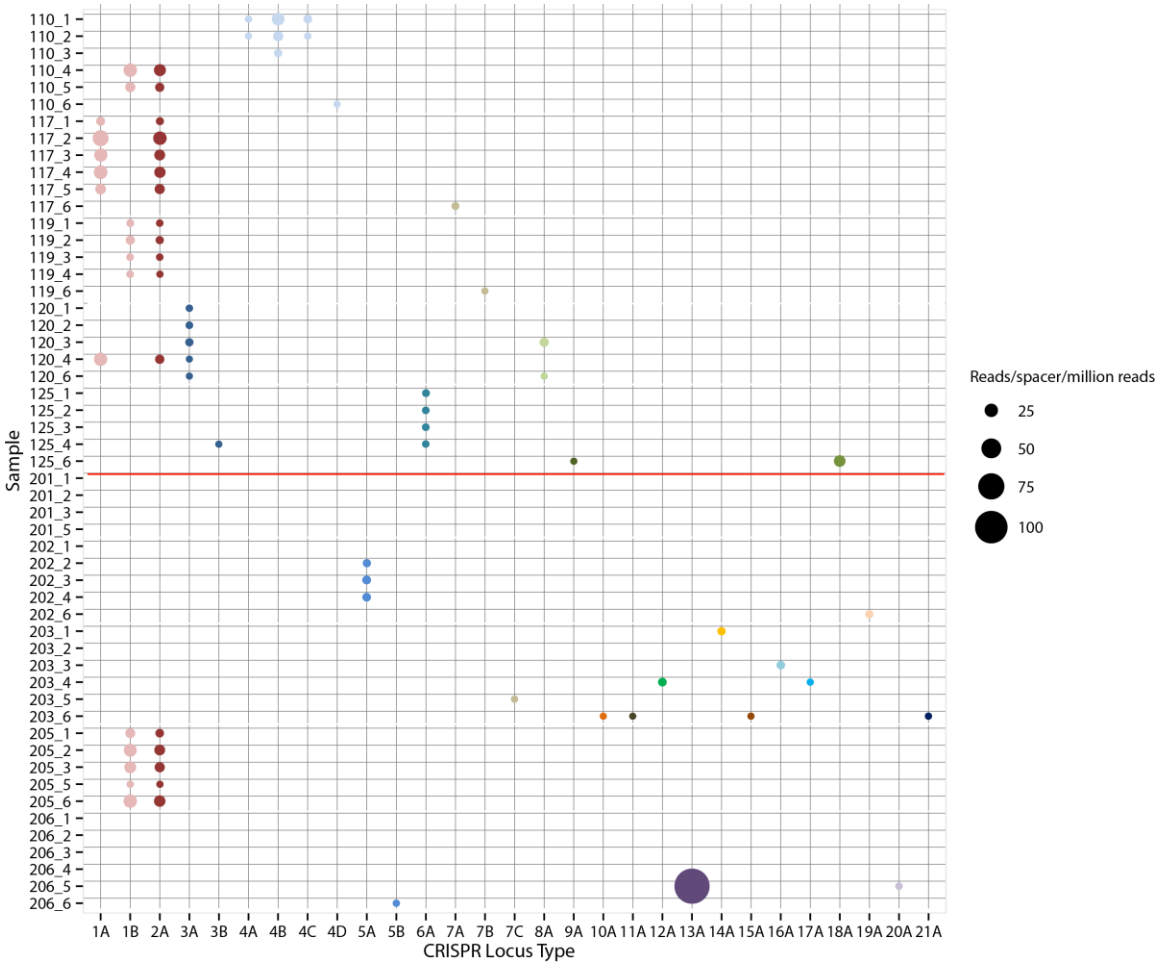


Figure 3.2. Spacer variation among subjects. Colored circles represent the number of reads matching each CRISPR locus type per spacer per million reads. The red line separates low-risk subjects (above) from high-risk subjects (below). Colors are consistent with those in Figure 3.1.

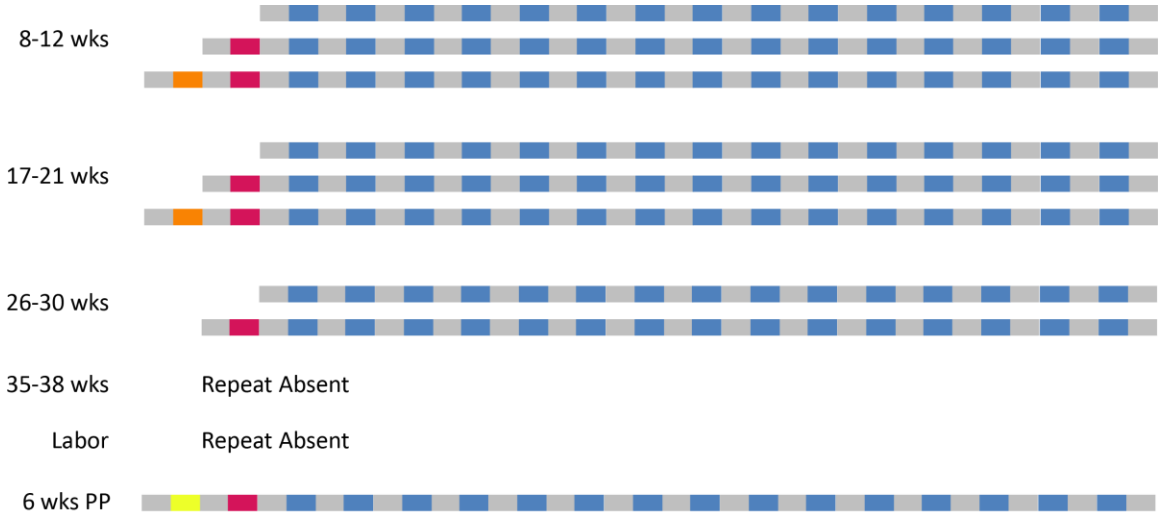


Figure 3.3. Spacer variation over time in the *L. iners* population of subject 110. Each line represents a repeat-spacer array found at the listed time point. Gray bars are repeats, blue bars are spacers shared between all alleles, and other colors represent spacers not present in all alleles. The leader end of the array is to the left. PP, postpartum.

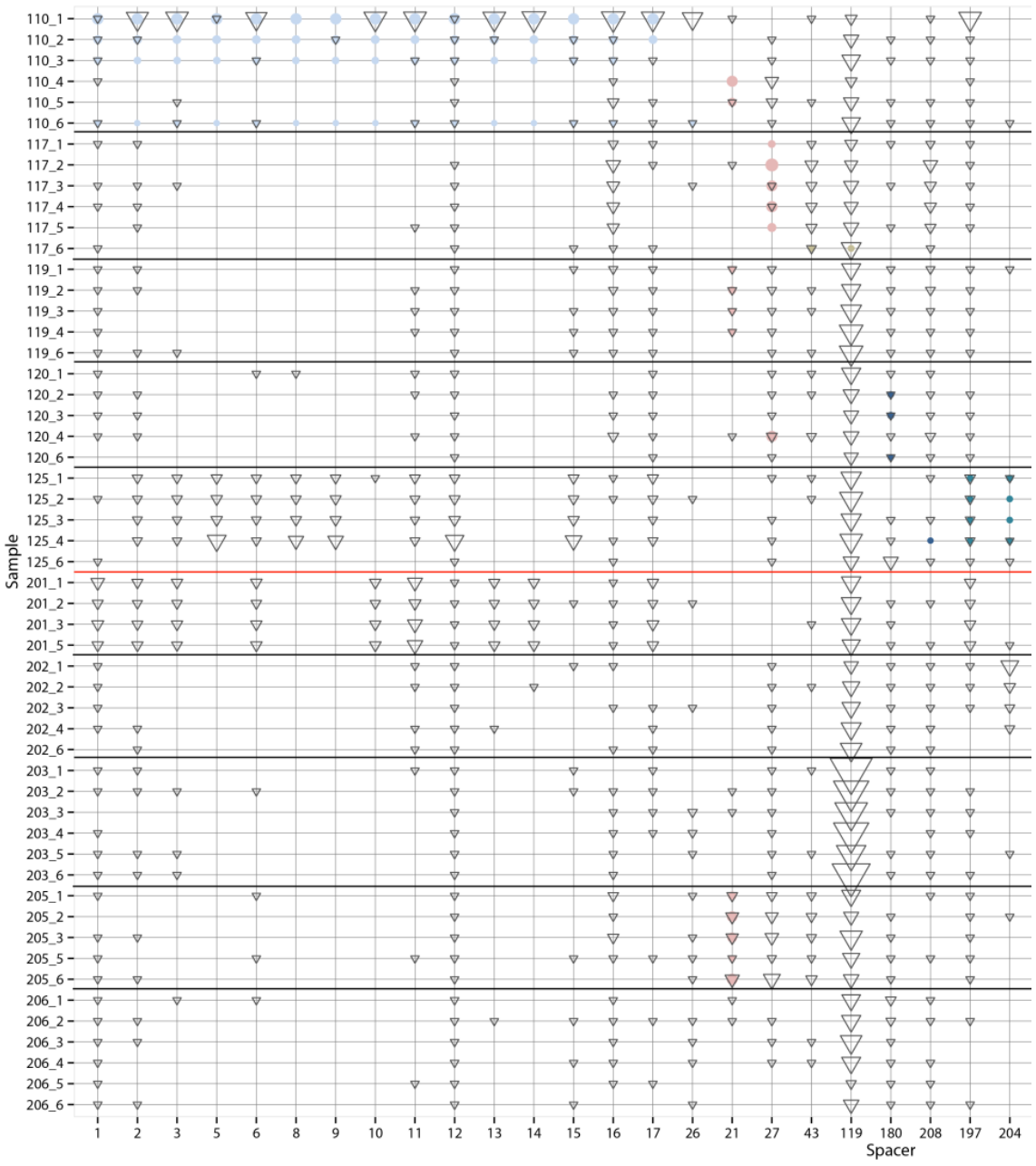


Figure 3.4. Spacer hits in the vaginal microbiome. Open triangles represent the number of BLAST hits from each spacer to the non-CRISPR portion of each sample, while colored circles represent the abundance of each spacer in the CRISPR loci of the sample an in Figure 3.2. Black lines separate individual subjects, while the red line separates low-risk subjects (above) from high-risk subjects (below).

Figure 3.4. (cont.)

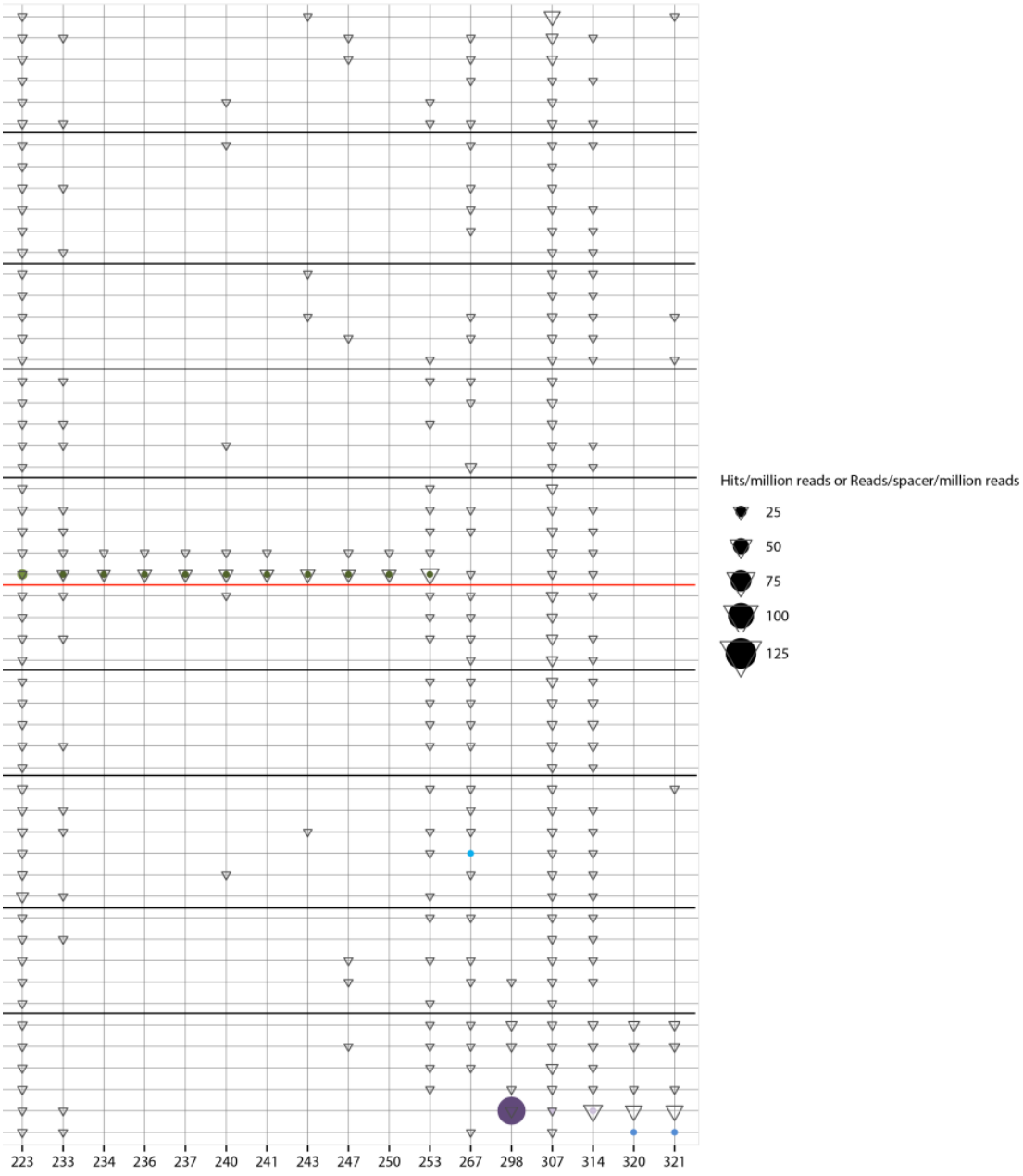


Table 3.1. Subjects and samples used in the study. PP: 6 weeks postpartum.

Subject ID	Preterm Risk Level	Samples	Subject ID	Preterm Risk Level	Samples
110	Low	8-12 wks 17-21 wks 26-30 wks 25-38 wks Labor PP	201	High	8-12 wks 17-21 wks 26-30 wks Labor
117	Low	8-12 wks 17-21 wks 26-30 wks 25-38 wks Labor PP	202	High	8-12 wks 17-21 wks 26-30 wks 25-38 wks PP
119	Low	8-12 wks 17-21 wks 26-30 wks 25-38 wks PP	203	High	8-12 wks 17-21 wks 26-30 wks 25-38 wks Labor PP
120	Low	8-12 wks 17-21 wks 26-30 wks 25-38 wks PP	205	High	8-12 wks 17-21 wks 26-30 wks Labor PP
125	Low	8-12 wks 17-21 wks 26-30 wks 25-38 wks PP	206	High	8-12 wks 17-21 wks 26-30 wks 25-38 wks Labor PP

References

[1] Aagaard, K., Riehle, K., et al. (2012). A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One 7* (6), e36466.

[2] Altschul, S.F., Gish, W., et al. (1990). Basic local alignment search tool. *J. Mol. Biol. 215* (3), 403–410.

[3] Barrangou, R., Fremaux, C., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science 315* (5819), 1709–1712.

[4] Baugher, J.L., Durmaz, E., et al. (2014). Spontaneously induced prophages in *Lactobacillus gasseri* contribute to horizontal gene transfer. *Appl. Environ. Microbiol. 80* (11), 3508–3517.

[5] Brouns, S.J.J., Jore, M.M., et al. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science 321* (5891), 960–964.

[6] Childs, L.M., England, W.E., et al. (2014). CRISPR-Induced distributed immunity in microbial populations. *PLoS ONE 9* (7), e101710.

[7] Childs, L.M., Held, N.L., et al. (2012). Multiscale model of CRISPR-induced coevolutionary dynamics: Diversification at the interface of Lamarck and Darwin. *Evolution 66* (7), 2015–2029.

[8] Costello, E.K., Lauber, C.L., et al. (2009). Bacterial community variation in human body habitats across space and time. *Science 326* (5960), 1694–1697.

[9] Cui, Y., Li, Y., et al. (2008). Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. *PLoS One 3* (7), e2652.

[10] Goldenberg, R.L., Culhane, J.F., et al. (2008). Epidemiology and causes of preterm birth. *Lancet 371* (9606), 75–84.

[11] Held, N.L., Herrera, A., et al. (2010). CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS One 5* (9), e12988.

[12] Held, N.L., Herrera, A., et al. (2013). Reassortment of CRISPR repeat-spacer loci in *Sulfolobus islandicus*. *Environ. Microbiol. 15* (11), 3065–3076.

[13] Horvath, P., Romero, D.A., et al. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol. 190* (4), 1401–1412.

[14] Hyman, R.W., Fukushima, M., et al. (2014). Diversity of the vaginal microbiome correlates with preterm birth. *Reprod. Sci. 21* (1), 32–40.

[15] Labrie, S.J., Samson, J.E., et al. (2010). Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol. 8* (5), 317–327.

[16] Makarova, K.S., Wolf, Y.I., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol. 13* (11), 722–736.

[17] Mastromarino, P., Di Pietro, M., et al. (2014). Effects of vaginal lactobacilli in *Chlamydia trachomatis* infection. *Int. J. Med. Microbiol. 304* (5–6), 654–661.

[18] Mojica, F.J.M., Díez-Villaseñor, C., et al. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol. 60* (2), 174–182.

[19] Petricevic, L., Domig, K.J., et al. (2014). Characterisation of the vaginal *Lactobacillus* microbiota associated with preterm delivery. *Sci. Rep. 4* , 5136.

[20] Pybus, V. and Onderdonk, A.B. (1999). Microbial interactions in the vaginal ecosystem, with emphasis on the pathogenesis of bacterial vaginosis. *Microbes Infect. 1* (4), 285–292.

[21] Ravel, J., Brotman, R.M., et al. (2013). Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. *Microbiome 1* (1), 29.

[22] Ravel, J., Gajer, P., et al. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U. S. A. 108 Suppl 1* , 4680–4687.

[23] Rho, M., Wu, Y.-W., et al. (2012). Diverse CRISPRs evolving in human microbiomes. *PLoS Genet. 8* (6), e1002441.

[24] Rice, P., Longden, I., et al. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet. 16* (6), 276–277.

[25] Riehm, J.M., Vergnaud, G., et al. (2012). *Yersinia pestis* lineages in Mongolia. *PLoS One 7* (2), e30624.

[26] Romero, R., Hassan, S.S., et al. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome 2* (1), 4.

[27] Skennerton, C.T., Imelfort, M., et al. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res. 41* (10), e105.

[28] Smoot, M.E., Ono, K., et al. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics 27* (3), 431–432.

[29] Srinivasan, S., Hoffman, N.G., et al. (2012). Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS One 7* (6), e37818.

[30] Turovskiy, Y., Sutyak Noll, K., et al. (2011). The aetiology of bacterial vaginosis. *J. Appl. Microbiol. 110* (5), 1105–1128.

[31] Tyson, G.W. and Banfield, J.F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol. 10* (1), 200–207.

[32] Walther-António, M.R.S., Jeraldo, P., et al. (2014). Pregnancy's stronghold on the vaginal microbiome. *PLoS One 9* (6), e98514.

[33] Wen, A., Srinivasan, U., et al. (2014). Selected vaginal bacteria and risk of preterm birth: an ecological perspective. *J. Infect. Dis. 209* (7), 1087–1094.

CHAPTER 4

Detection of CRISPR Spacers in Metagenomes from the Cystic Fibrosis Lung

Abstract

Viruses can have dramatic impacts on microbial communities, and in the human microbiome, virus-microbe interactions can influence the health of the human host. The microbial community of the cystic fibrosis lung plays a critical role in the clinical outcome of the disease, and *Pseudomonas aeruginosa* infection in particular is responsible for much of the morbidity and mortality associated with this condition. The CRISPR immune system is an important mechanism by which microbes gain immunity to viruses, and diversity of spacer content may affect population stability. To better understand how CRISPR diversity influences how *P. aeruginosa* and other bacteria interact with phages, we sequenced bacterial and viral metagenomes from sputum and explant lung samples from 12 cystic fibrosis patients and identified CRISPR sequences in a focal subset of samples. We found highly varied *P. aeruginosa* spacer content between patients, but no evidence of within-patient diversity despite deep sequencing. Many of the identified spacers matched known phages or other foreign genetic elements; however, few of these elements were found in the samples containing hosts with matching spacers. These findings raise questions about the phage-host interactions in the cystic fibrosis lung and open the door to future investigation of other bacteria which colonize this environment.

Introduction

The microbial communities which comprise the human microbiome are no exception to the effects of microbe-virus interactions. Viruses have been observed all over the human body, and viral communities have been studied in the oral cavity, on the skin, in the bloodstream, in the respiratory tract, and in the gut [1]. Direct effects on microbial communities by phage have also been observed; in a gnotobiotic mouse model of the human gut microbiota, the gut community was subjected to a phage attack, resulting in shifts in microbial community dynamics [21].

Among the many human-associated microbial communities which impact human health and disease is the community of the cystic fibrosis lung. Cystic fibrosis is among the most common life-shortening human genetic disorders, particularly in those of northern European descent, among whom the disease is present in 1 in 3,000 live births [20]. While the disease is genetic in origin, the vast majority of morbidity and mortality stems from persistent microbial infections in the lung [6]. Among the most common and damaging of these pathogens is *Pseudomonas aeruginosa*, which plays a key role in cystic fibrosis morbidity and mortality [13]. *P. aeruginosa* is a ubiquitous opportunistic pathogen, and cystic fibrosis patients are believed to primarily be colonized by strains encountered in their environment, which subsequently adapt to the lung environment and thrive [13]. Once established, *P. aeruginosa* infections are difficult to eradicate, owing to biofilm formation and widespread antibiotic resistance [7].

The CRISPR adaptive immune system is one method by which microbial hosts can gain immunity to viruses, and CRISPRs could have significant impacts in microbial communities on multiple levels. Their ability to target many types of foreign and mobile genetic elements contributes to viral resistance while also restricting horizontal gene transfer, leading to a tradeoff between the potential benefits of introducing new genes and protection against viral lysis or other negative consequences. Though some systems have been found to strike a balance by tolerating temperate phage integration while preventing lytic induction [10], in many other cases, active CRISPR systems prevent any maintenance of the matched element in the cell [3,8,18]. As a result of their ability to add a wide array of potential spacers to any given element, CRISPR arrays are among the most diverse regions of the microbial genome. Diversity of spacer content may also have an effect on population stability through distribution of immunity; simulated populations with varied spacers to a similar phage pool were found to be more stable, with corresponding viral populations more prone to extinction [5].

To investigate diversity in the CRISPR system of the microbes which inhabit the cystic fibrosis lung, we sequenced bacterial and viral metagenomes from cystic fibrosis sputum and lung explant samples and identified CRISPR alleles from all possible species. *P. aeruginosa* was chosen as a representative species to describe its CRISPR diversity,

spacer targets, and the dynamics of protospacer-containing elements in matched viral metagenomes.

Results

Bacterial species presence and abundance varies among cystic fibrosis patients

For our initial screen, 10 total sputum samples were obtained from five patients at one to three time points (Figure 4.1) and sequenced. Taxonomic classification of the resulting reads revealed major differences in the composition of these communities (Figure 4.2). Samples were clearly divided into two major categories by the dominant organism. Patients CF15 and CF16 were dominated by *Rothia mucilaginosa*, a common resident of the human oral cavity and upper respiratory tract which can become an opportunistic pathogen in cystic fibrosis patients [17]. Samples CF15B and CF16A were most strongly dominated by *R. mucilaginosa*, with over 80% of reads in each sample coming from this organism. Sample CF16B was 56% *R. mucilaginosa*, with an additional 28% of the reads belonging to *Enterococcus faecalis*, a species not found in any of our other four patients. Sample CF15A was the richest and most even of the samples analyzed; while dominated by *R. mucilaginosa* at 25% of reads, it contained two other species making up 10% or more of the total read abundance: *Streptococcus parasanguinis* (14%), *P. aeruginosa* (11%), as well as numerous other species making up smaller proportions of the community.

By contrast, patients CF14, CF17, and CF19 were dominated by *Staphylococcus aureus*, often one of the first pathogens to colonize the lungs of children with cystic fibrosis [11]. All but one of the six samples from these patients was dominated by *S. aureus*, with a relative abundance of 60% and even reaching as high as 94% in one sample, with the remaining reads largely belonging to *Streptococcus* and *Rothia* species. The sole exception is sample CF14C, which contained only 10% *S. aureus* and was instead dominated by *Escherichia* (56% *E. coli* and 14% *Escherichia* reads which could not be classified at the species level).

Identification of CRISPR spacers in metagenomes

We next conducted a broad survey of CRISPRs present in all species in these metagenomes. To capture the maximum number of spacers possible, even those without known repeat sequences, we used a pipeline including Crass [27], a program capable of detecting CRISPR-like sequences *de novo* in metagenomic sequence reads, to identify potential CRISPR loci in these metagenomes (see Methods for details). We chose to focus on patients CF15A and CF16A, as the others were strongly dominated by *S. aureus*, which rarely harbors CRISPR loci [12]. Specifically, we chose samples CF15A and CF16A for their relative species-level diversity, and because many of the species present were known to contain CRISPRs [12]. We identified 17 CRISPR repeats in sample CF15A and four in the less diverse CF16A, for a total of 20 unique repeats (Figure 4.3). In CF15A, the most prevalent repeat matches a repeat found in *Rothia* genomes; however, despite the fact that both samples are dominated by *Rothia*, we failed to identify any *Rothia*-linked repeats in sample CF16A. This could be the result of different strains of *Rothia* residing in these two patients, only one of which contains a CRISPR system. In keeping with CF15A having a greater species richness, it contains far more different repeat types belonging to different species than CF16A. In fact, the only repeat type shared between these two samples is one belonging to *P. aeruginosa*.

P. aeruginosa *CRISPRs lack coexisting diversity in metagenome samples*

Given the presence of *P. aeruginosa* CRISPR repeats in both of our focal samples, and the status of the species as an important cystic fibrosis pathogen, we chose to investigate the spacer content and diversity of *P. aeruginosa* CRISPRs in these two metagenomes in more detail. We used a network analysis-based method (see Methods) to identify repeat-spacer arrays as well as variant loci. *P. aeruginosa* strains typically possess Type I CRISPR loci, most commonly of subtype I-F and less frequently of subtype I-E; both types may be present in the same strain, and can be differentiated by repeat sequence or types of *cas* genes present [4]. We observed two repeat-spacer arrays in each of these samples with repeat sequences associated with Type I-F CRISPRs, containing 12 to 18 spacers each, with no evidence of variant alleles (Figure 4.4). *P. aeruginosa* CRISPR systems commonly contain two repeat-spacer arrays with very similar repeats, so while

the two arrays cannot be conclusively linked to each other, it is likely they represent two parts of a single CRISPR system, so the lack of variation observed indicates a single CRISPR-type is present in each sample. The arrays from the two samples are completely distinct, sharing no spacers with one another.

To broaden our sample size of *P. aeruginosa* CRISPRs, we also identified CRISPRs in sample CF15B, the only other sample with detectable *P. aeruginosa*; however, coverage was not sufficient to fully assemble CRISPR loci, and all identified spacers were also present in CF15A, leading us to conclude that CF15B contains the same CRISPR-type as CF15A (Figure 4.4). We also sequenced metagenomes of four additional samples: two sputum samples from consecutive days from a longitudinal study of a single cystic fibrosis patient, and two samples from a cystic fibrosis lung explant (Figure 4.1). Despite all samples being high in *P. aeruginosa* (Figure 4.2), we again found two Type I-F arrays, with no variation in the explant samples; the same pair of arrays was found in both samples. In addition, we found two additional shorter arrays with a different repeat type associated with Type I-E CRISPRs (Figure 4.4). As before, all the spacers in these arrays were unique and not found in the other samples. In the longitudinal samples, no evidence of *P. aeruginosa* CRISPRs was found.

Identified CRISPRs are similar to previously sequenced CRISPRs

To determine if the CRISPR loci extracted from our metagenomes were related to those from any previously sequenced *P. aeruginosa* strains, we compared all of our identified spacers to the NCBI nt database using blastn. We found that the CF15A CRISPR-type shared all of its spacers with the small colony variant isolate SCV20265, isolated from a cystic fibrosis patient in Hannover, Germany, though the SCV strain had two additional spacers in one array that did not appear in our metagenomes (Figure 4.5). The CF16A arrays were identical to those found in SMC4501, a sputum isolate from 1999 in Hanover, New Hampshire (Figure 4.5). The explant strain held CRISPRs substantially similar to those in SMC4489, a *P. aeruginosa* strain isolated from an eye infection in Pittsburgh, Pennsylvania in 1994; all of the Type I-F and I-E spacers we found were present in the eye isolate; however, that isolate had an additional Type I-E array containing spacers not found in our metagenome samples (Figure 4.5).

Identification of matched protospacer sequences

Spacers identified in the three metagenomic samples were blasted against the NCBI nt database to identify known viruses, plasmids, or other elements they matched. In CF15A, 15/30 spacers (50%) matched a known phage, prophage, plasmid, or genomic island (Figure 4.6). By contrast, only 7/28 spacers (25%) found in sample CF16A had known matches; intriguingly, one of the two arrays contained no spacers with known matches (Figure 4.6). In the Type I-F arrays of the explant samples, 17/30 spacers (57%) matched known elements; the Type I-E arrays had 8/11 spacers (73%) with matches (Figure 4.6). Most spacers (34) matched known phages, with smaller numbers matching plasmids (3) and non-phage genomic islands (7). Sixteen also matched prophage characterized in the Liverpool epidemic strain LESB58 of *P. aeruginosa*. Strikingly, every spacer with a known match in the Type I-E arrays found in explant samples matched three such prophage, while in the Type I-F systems of all three samples, only two to three spacers matched these prophage (Figure 4.6).

Protospacer presence in host and viral metagenomes

To determine if any of the elements matched by our identified spacers were present in the corresponding sample, we used blastn to compare the spacers to both the host metagenomes (excluding reads identified as being part of CRISPR loci) and matched viral metagenomes for samples CF15A and CF16A (a matched viral sample was unavailable for the explant samples). We found two spacers from CF15A matched protospacers in the CF15A viral metagenome; one matching phiCTX and one matching phages D3, phi297, and vB_PaeS_PMG. In the host metagenome, we found four matched protospacers at modest to low coverage: the same D3-matching spacer observed in the viral metagenome, plus spacers matching the pathogenicity island PAPI-1 and phages JBD18 and D3112 (Figures 4.6 & 4.7). In the CF16A viral metagenome, we found five matched protospacers, all at relatively low abundance: one spacer matching the genomic island PAGI-6, one matching phage F10, one matching many phage including DMS3, D3112, and LES prophage 4, and two matching no known phages, plasmids, or genomic islands. In the host metagenome, we found three low-abundance protospacers, two of which had no known matches and one of which matches phages D3

and vB_PaeS_PMG. None of these overlap with the spacers found in the viral sample, and all are at a lower relative abundance than the average spacer (Figures 4.6 & 4.7).

In contrast to the general low number of matches and low abundance in the matched metagenomes, we found a protospacer at higher abundance in our CF15A host and viral metagenomes which was matched by a CF16A protospacer (Figure 4.7). This spacer, which also appears in the CF16A viral metagenome at a much lower relative abundance, matches several phages, including DMS3, D3112, and LES prophage 4 (Figure 4.6). This was by far the most abundant protospacer identified, whether in sample from the same patient or across samples.

Discussion

In our investigation of CRISPR diversity in the cystic fibrosis lung metagenome, we found an abundance of CRISPR-containing species, including the critical cystic fibrosis pathogen *P. aeruginosa*. In samples from three patients, we found completely non-shared CRISPR spacers, highlighting the diversity of spacers among *P. aeruginosa* strains; however, each of these CRISPR-types was substantially similar to a previously sequenced strain from distant locations, showing that related strains have migrated globally. Despite the variation between patients, we observed no CRISPR diversity within any of the patients. The spacers we observed in these three distinct CRISPR-types match many known phages and prophages as well as some plasmids and genomic islands, but over half of the spacers we found match no known elements. Few protospacers matched by these spacers are present in the same host metagenome, or in matched viral samples, though we do detect matching protospacers in other metagenomes. These findings reveal low levels of intra-patient diversity coupled to a globally-distributed, diverse *P. aeruginosa* population and incite further investigation into CRISPR diversity in *P. aeruginosa* and other species in the cystic fibrosis lung environment.

At the outset of this study, we hypothesized there would be coexisting CRISPR spacer alleles within a single patient at the same time point based on previous work showing that diverse immunity encourages microbial population stability by preventing viruses from evading all host immunity through a single escape mutation [5]. However, our close

investigation of *P. aeruginosa* CRISPRs revealed that this was not the case. Why might the cystic fibrosis lung environment not foster local diversity in *P. aeruginosa* CRISPRs? It is possible that *P. aeruginosa* does not experience significant phage pressure in this environment, which would both limit its ability to diversify (by restricting the pool of elements from which the CRISPR system can acquire new spacers) and remove the fitness advantage the host microbe gains from such diversity. However, lytic *P. aeruginosa* phage activity has been shown in chronic cystic fibrosis lung infections [14]. Though this study only investigated a specific subset of temperate phages present in LES strains and may not be reflective of the dynamics of other *P. aeruginosa* phage, it is still clear that phages play an important role in cystic fibrosis lung ecology. It is also possible that *P. aeruginosa* is employing other resistance mechanisms to deter phage, making CRISPR diversity less advantageous. Further sampling of cystic fibrosis patients is needed to determine if this lack of diversity is generally representative of *P. aeruginosa* in the cystic fibrosis lung.

Despite the lack of diversity within each patient, we observed completely different spacer arrays between patients, each of which was substantially similar to CRISPRs in other *P. aeruginosa* strains. These strains came from geographically distant sites and from different body sites – only one was from a cystic fibrosis patient. This observation indicates a ubiquitous *P. aeruginosa* population where strains can colonize varied environments, rather than separating into more specific niches. This is consistent with previous observations of a common *P. aeruginosa* clone type, clone C, in aquatic environments, cystic fibrosis clinics, and cystic fibrosis patients [23,24]. Other studies have also identified the environment rather than other patients as the primary source for acquisition in patients [29] and found that cystic fibrosis isolates are a random subset of the larger environmental population [15]. While certain transmissible strains such as Liverpool, Manchester, and Australian epidemic strains have adapted to infect the CF lung [9], the strains we found in these three patients appear to be part of a globally distributed population of *P. aeruginosa* which can move between environments.

Among our three identified CRISPR-types, we found spacers matching three integrated LES prophage in every array, save for one in CF16A. In the Type I-E arrays of the

explant samples, seven out of eleven spacers matched these prophages. The LES prophage have previously been found to confer a competitive advantage in a rat chronic lung infection model [32], and have been shown to exist as free phage with lytic activity in the human cystic fibrosis lung [14]. These data support a role for such prophage in competition between *P. aeruginosa* strains in the lung; our observation of numerous spacers matching these prophages shows that these strains have encountered these prophage or their close relatives repeatedly through time. Due to the Type I CRISPR system's protospacer adjacent motif (PAM)-based self-recognition system [19], any host with an active Type I system and one or more spacers matching a phage cannot integrate that phage into its genome; its Cas machinery will target the integrated prophage, leading to destruction of its own DNA. Therefore, possessing these spacers prevents the strains we found from incorporating these advantageous prophages, unless the CRISPR system has been inactivated. Consistent with this, we found that while the CF15A and CF16A strains both had spacers matching LES prophage, few to no reads containing those protospacers were found in their respective host metagenome. One LES protospacer matched by CF16A was found at higher abundance in the CF15A host metagenome sample, which lacks that spacer; however, that protospacer sequence is conserved in eight related sequenced phage (see Figure 4.6), and thus is not necessarily indicative of LES phage presence. Such a tradeoff reflects a delicate balance between the safety of CRISPR immunity to phages and other elements and the potential advantages these elements can bring, such as antibiotic resistance or pathogenicity islands. These strains appear to have taken the safe route, maintaining their ability to fight of invasions by LES strains; a comparison of their fitness to strains with missing or nonfunctional CRISPRs would be informative and may help further elucidate these tradeoffs.

While little diversity was observed in *P. aeruginosa* CRISPRs, there remains considerable untapped data in these metagenomes. In our preliminary scan of this data, we found 19 other CRISPR repeat types found in our two focal samples, and we anticipate finding even more in the remainder of these samples, especially those with different compositions. These repeats belong to at least 15 different genera, each with the potential to have its own level of CRISPR variation and distinct pattern of interactions with phage. Some of these genera are deeply covered in the sequence data; for example,

over 2,300 reads contain the *Rothia* repeat found in sample CF15A. Such samples offer great promise for identifying CRISPR diversity if it exists in these samples, as deeper coverage enables easier detection of rarer CRISPR variants. The potential to uncover more information about CRISPR diversity in the cystic fibrosis lung is vast, and investigating these additional taxa remains a promising area for future study.

Methods

Sample selection and processing

Samples were collected and processed by Douglas Conrad of the University of California San Diego and Yan Wei Lim of San Diego State University, with assistance from other members of the Forest Rohwer lab at San Diego State University. All samples were obtained from adult male cystic fibrosis patients at the University of California San Diego Adult Cystic Fibrosis Clinic, La Jolla, CA, during regular clinic visits, in accordance with the University of California and San Diego State University institutional review boards. Sample collection, processing, and DNA extraction were performed as in [16]. Briefly, induced sputum samples were collected following an inhalation of saline solution and an oral rinse to minimize contamination from oral microbiota. Samples were homogenized and aliquoted for separate protocols for bacterial and viral metagenome sequencing. Bacterial metagenome samples were treated with beta-mercaptoethanol to disrupt mucus, repeatedly washed to induce hypotonic lysis of human cells, and treated with DNase to remove DNA from lysed cells as well as extracellular microbial DNA from biofilm formation. Viral samples were treated with dithiothreitol to disrupt mucus, filtered, and subjected to CsCl gradient ultracentrifugation to isolate viral particles. DNA was extracted from both sample types, via CTAB/phenol-chloroform extraction for the viral samples and NucleoSpin tissue kit for bacterial samples. Lung explant samples were collected and processed as described in [31]. Briefly, the interior surface was cut from 2 cm x 4 cm lung tissue samples and homogenized before DNA extraction using a NucleoSpin tissue kit.

Metagenome sequencing

All bacterial and viral metagenomes were prepared and sequenced by the Roy J. Carver Biotechnology Center at the University of Illinois. Libraries were prepared using TruSeq Sample Prep kits (Illumina, San Diego, CA). Bacterial samples from patients CF14, CF15, CF16, CF17, and CF19 were each sequenced using a full lane on an Illumina MiSeq v2 sequencer (Illumina, San Diego, CA), producing 250 nt paired-end reads. Focal bacterial samples CF15A and CF16A were additionally sequenced using a full lane per sample on an Illumina HiSeq2500 (Illumina, San Diego, CA), producing 160 nt paired-end reads. Viral samples CF15A and CF16A were each sequenced using a full lane of Illumina MiSeq, as with the bacterial samples. Explant and longitudinal samples were sequenced using an Illumina HiSeq2500. One sample, Explant C, received a full lane of sequencing, while the remaining explant sample and two longitudinal samples were barcoded and pooled in a second lane. Sequencing adapters were trimmed from the resulting reads by the sequencing center.

Quality filtering and human sequence removal

All sequenced samples were quality filtered using Prinseq 0.20.4 [26]. Reads were dereplicated to eliminate PCR duplicates and trimmed from the left and right ends using a 5nt sliding window with a minimum quality score of 30. Reads were retained if they had a mean quality score of 30 and less than 1% ambiguous bases. The minimum read length was set to approximately two-thirds the anticipated read length, resulting in a cutoff of 100 nt for HiSeq data and 160 nt for MiSeq data. After quality filtering, contaminating human sequence was removed using DeconSeq [25]; reads matching databases of bacterial or viral sequences and those matching no databases were retained, while those matching human sequence databases were removed.

Estimation of relative abundance

In all bacterial samples, relative abundance of taxa was calculated using MetaPhlAn2 [30], which classifies reads using clade-specific marker genes from 17,000 reference genomes to provide a species-level estimate of relative abundance. Default parameters were used.

CRISPR identification

For focal samples CF15A and CF16A, potential CRISPR repeat sequences in each sample were identified using CRASS [27]. Potential repeats were compared via blastn [2] to the NCBI nt database; any sequences that primarily hit non-microbial sequences were classified as false positives and excluded from further analysis. Reads matching remaining potential CRISPR repeats were identified using blastn [2] (blastn-short, e<=0.1, >=90% nucleotide identity). To establish spacer order and detect coexisting CRISPR alleles, spacers were linked as follows: 12-nucleotide "chunks" of DNA flanking each repeat were identified using fuzznuc [22]; up to 8 mismatches to the repeat sequence were allowed to capture degenerate repeats. Due to the palindromic nature of the repeats, this occasionally generated situations where the repeat was matched in both orientations; in these cases, the match with fewer mismatches was kept and the secondary match discarded. Chunks that were a perfect match to the repeat sequence (i.e., from two adjacent repeats or partial repeats) were also discarded. Finally, singleton chunks that perfectly overlap non-singleton chunks by at least 8 bp were removed, to account for rare chunks generated by sequencing error.

Custom bash scripts were used to identify occurrences of two or more chunks on the same read. These were recorded as links, which represent either two ends of the same spacer, or opposite ends of two spacers linked across a repeat. The first type of link was used to identify spacer sequences; the second, to order spacers and identify branch points, i.e. where multiple alleles with partially shared spacer content exist. Linkage networks were analyzed using Cytoscape [28]. Based on average repeat and spacer lengths of species with previously sequenced CRISPR loci, links spanning a single repeat-spacer unit were considered short links, spanning only a single spacer or pair of adjacent spacers; longer links were considered to span multiple spacers and were not counted when determining coverage of links across a spacer or between adjacent spacers.

Comparison of identified spacers to sequenced P. aeruginosa *and extrachromosomal element sequences*

All spacers from our three identified CRISPR-types were blasted against the NCBI nt database using blastn (blast-short, e= 0.1). For matches to other CRISPR arrays, only full-length perfect matches to *P. aeruginosa* strains were considered, and regions flanking the hits were examined to ensure they contained CRISPR repeat sequence. For matches to phages and plasmids, matches with an e-value below 0.1 and at least 90% coverage of the query were retained.

Identification of matched protospacers in metagenomes

All spacers identified in CF15A and CF16A were compared to local blast databases of reads from host and viral metagenomes from those samples using blastn (blastn-short, e=0.1). Reads previously identified as containing CRISPR repeats were excluded from analysis. The number of hits was normalized by the number of reads in the database.

Figures

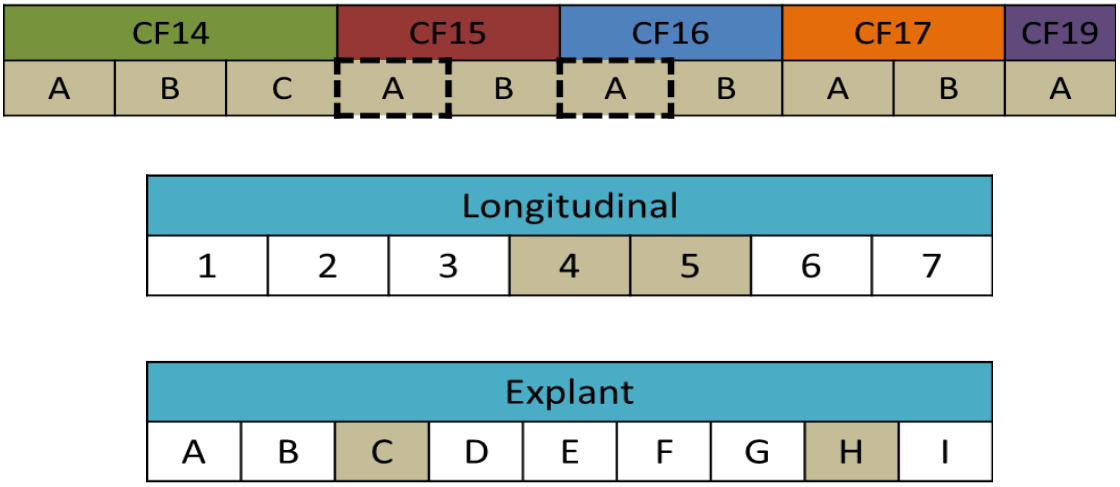


Figure 4.1. Samples used in this study. The upper bar represents the patient; the lower boxes represent individual samples taken at different times (CFxx & Longitudinal) or in different areas of the lung (Explant). Samples shaded in tan have host metagenome sequences; those with bold dotted outlines additionally have matched viral metagenomes sequenced.

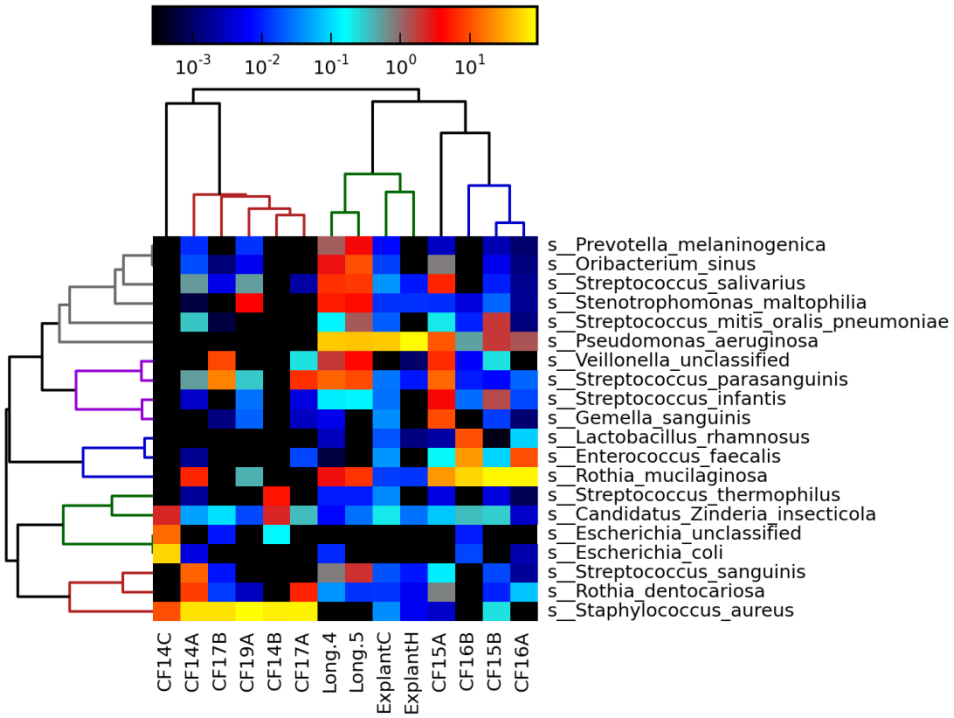


Figure 4.2. Heatmap of taxonomic classifications of metagenome reads. Reads were assigned to species-level taxa using MetaPhlAn [30]. Colors indicate higher or lower abundance of species in accordance with the scale at the top of the figure; black boxes represent absent taxa. Species are grouped vertically per the tree to the left; samples are grouped horizontally by shared taxa as shown by the tree at the top of the figure.

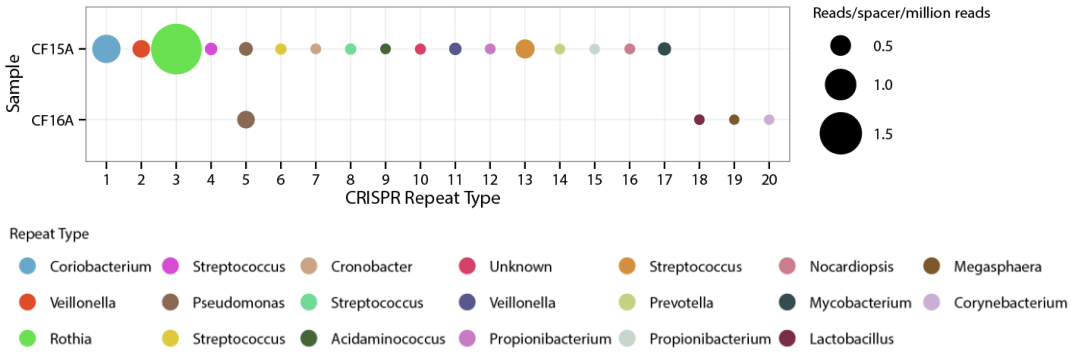


Figure 4.3. CRISPR repeats identified in two metagenomes. Each circle represents a repeat type. Area of the dot is scaled to the number of reads containing the repeat, normalized to the number of spacers in the repeat-spacer array and the number of reads in the sample. Repeat types are assigned to the most likely organism at the genus level based on repeats identified in publically available sequence data [12].

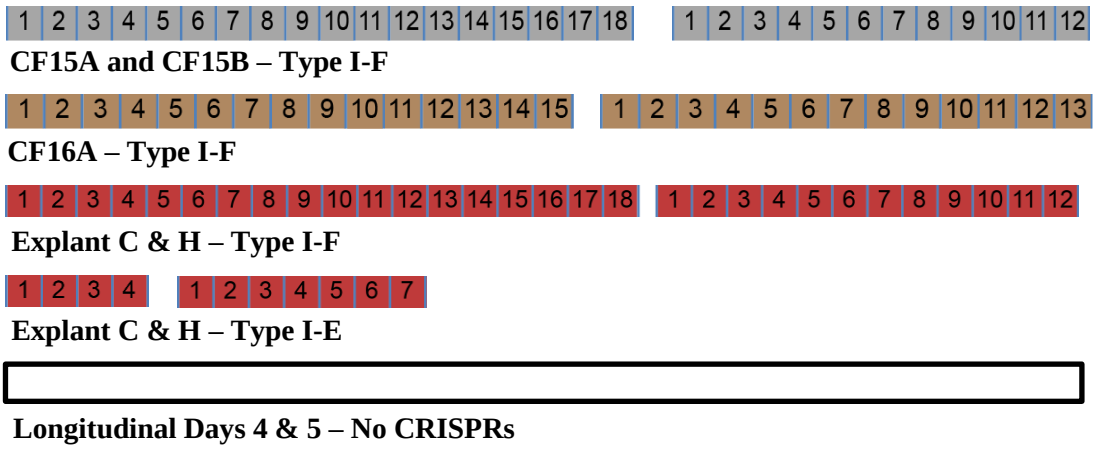


Figure 4.4. *P. aeruginosa* repeat spacer arrays identified in metagenomes. Each numbered block is a spacer; each contiguous block of spacers represents an individual array. Arrays of the same type from the same sample are grouped horizontally.

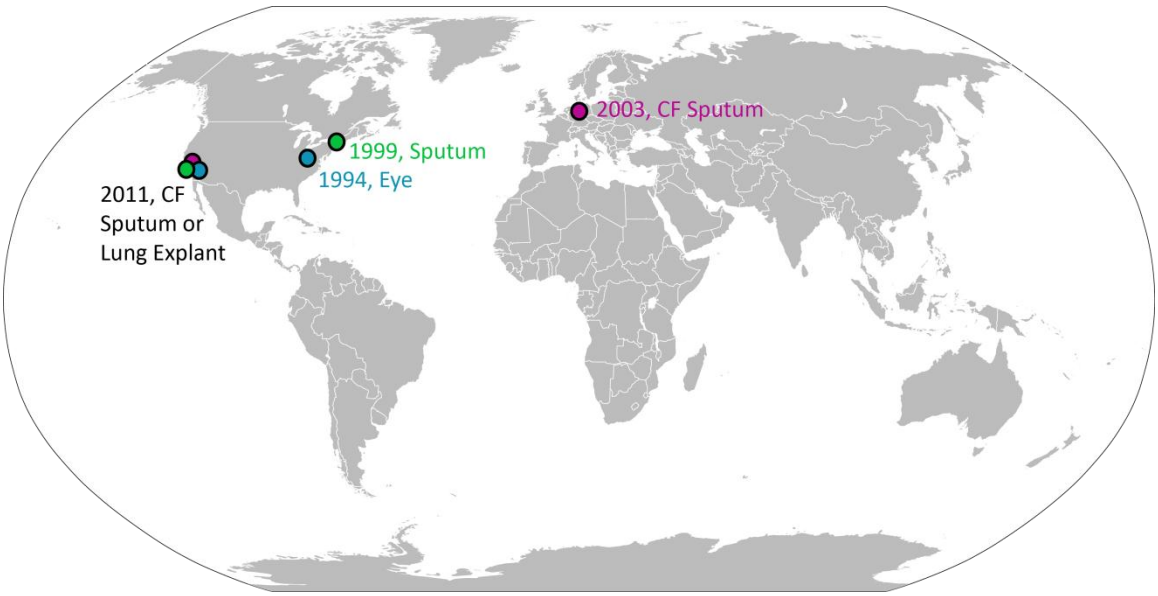


Figure 4.5. *P. aeruginosa* from San Diego cystic fibrosis patients have similar CRISPRs to globally distributed strains. Each colored circle represents a CRISPR-type identified from our metagenomes or from publicly available sequence data. Color-matched circles share all or most of their spacers. Sampling date and body site are shown for each sample.

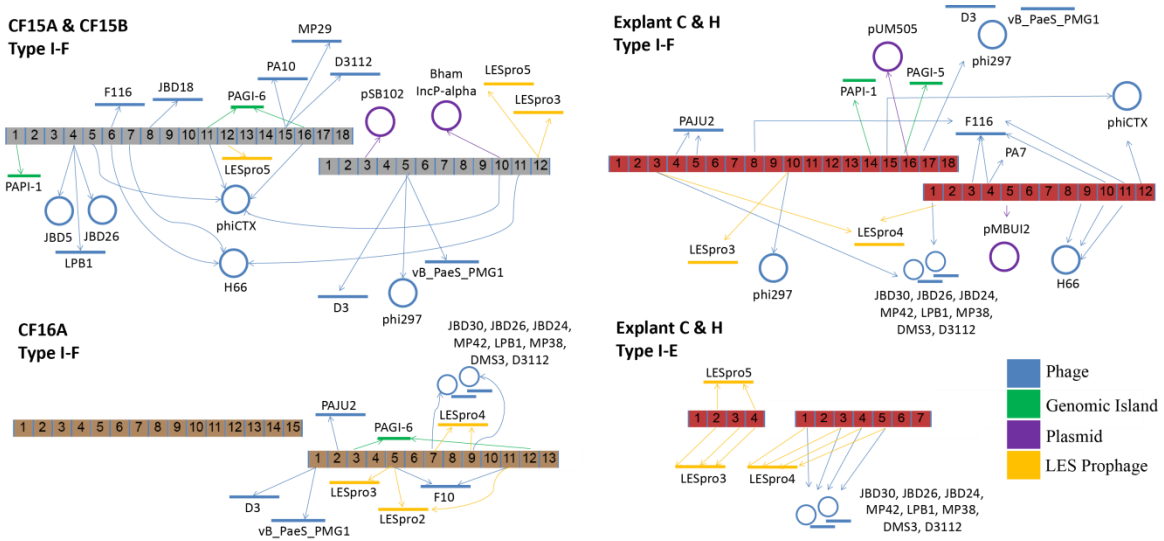


Figure 4.6. ***P. aeruginosa*** **spacers match known extrachromosomal elements.** Arrays are as in Figure 4.4. Matched elements are shown as linear or circular according to their DNA structure and color-coded by element type. An arrow from a spacer to an element indicates a spacer match.

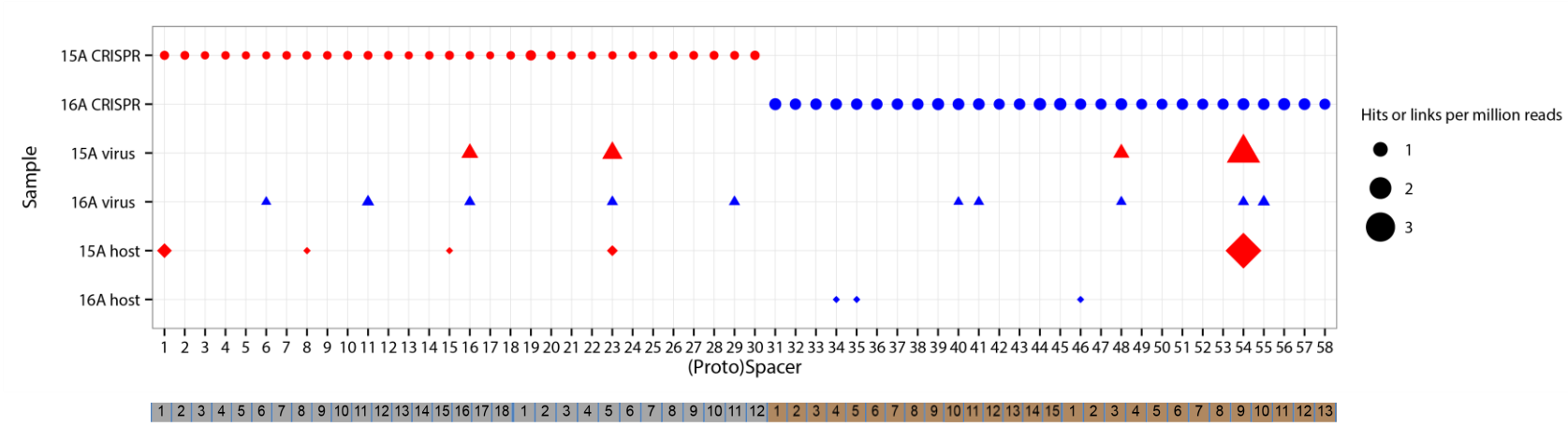


Figure 4.7. *P. aeruginosa* protospacers in host and viral metagenomes. For each spacer, hits to CRISPRs in the host metagenome (circles), viral metagenome (triangle), and non-CRISPR regions of the host metagenome (diamonds) are shown scaled to their relative abundance in that metagenome. Matches to the CF15A metagenomes are shown in red; CF16A metagenomes are in blue. Arrays are as in Figure 4.5 and 4.6, with array maps shown at the bottom for reference.

References

[1] Abeles, S.R. and Pride, D.T. (2014). Molecular bases and role of viruses in the human microbiome. *J. Mol. Biol. 426* (23), 3892–3906.

[2] Altschul, S.F., Gish, W., et al. (1990). Basic local alignment search tool. *J. Mol. Biol. 215* (3), 403–410.

[3] Bikard, D., Hatoum-Aslan, A., et al. (2012). CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe 12* , 177–186.

[4] Cady, K.C., White, A.S., et al. (2011). Prevalence, conservation and functional analysis of *Yersinia* and *Escherichia* CRISPR regions in clinical *Pseudomonas aeruginosa* isolates. *Microbiology 157* (2), 430–437.

[5] Childs, L.M., England, W.E., et al. (2014). CRISPR-induced distributed immunity in microbial populations. *PLoS One 9* (7), e101710.

[6] Cystic Fibrosis Foundation. (2015). *Patient Registry Annual Data Report 2014*.

[7] Drenkard, E. and Ausubel, F.M. (2002). *Pseudomonas* biofilm formation and antibiotic resistance are linked to phenotypic variation. *Nature 416* (6882), 740–743.

[8] Edgar, R. and Qimron, U. (2010). The *Escherichia coli* CRISPR system protects from λ lysogenization, lysogens, and prophage induction. *J. Bacteriol. 192* (23), 6291–6294.

[9] Fothergill, J.L., Walshaw, M.J., et al. (2012). Transmissible strains of *Pseudomonas aeruginosa* in cystic fibrosis lung infections. *Eur. Respir. J. 40* (1), 227–238.

[10] Goldberg, G.W., Jiang, W., et al. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature 514* (7524), 633–637.

[11] Goss, C.H. and Muhlebach, M.S. (2011). Review: *Staphylococcus aureus* and MRSA in cystic fibrosis. *J. Cyst. Fibros. 10* (5), 298–306.

[12] Grissa, I., Vergnaud, G., et al. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics 8* , 172.

[13] Hauser, A.R., Jain, M., et al. (2011). Clinical significance of microbial infection and adaptation in cystic fibrosis. *Clin. Microbiol. Rev. 24* (1), 29–70.

[14] James, C.E., Davies, E.V., et al. (2015). Lytic activity by temperate phages of *Pseudomonas aeruginosa* in long-term cystic fibrosis chronic lung infections. *ISME J. 9* (6), 1391–1398.

[15] Kidd, T.J., Ritchie, S.R., et al. (2012). *Pseudomonas aeruginosa* exhibits frequent recombination, but only a limited association between genotype and ecological setting. *PLoS One 7* (9), e44199.

[16] Lim, Y.W., Evangelista, J.S., et al. (2014). Clinical insights from metagenomic analysis of sputum samples from patients with cystic fibrosis. *J. Clin. Microbiol. 52* (2), 425–437.

[17] Lim, Y.W., Schmieder, R., et al. (2013). Mechanistic model of *Rothia mucilaginosa* adaptation toward persistence in the CF lung, based on a genome reconstructed from metagenomic data. *PLoS One 8* (5), e64285.

[18] Marraffini, L.A. and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science 322* (5909), 1843–1845.

[19] Mojica, F.J.M., Díez-Villaseñor, C., et al. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology 155* (3), 733–740.

[20] O'Sullivan, B.P. and Freedman, S.D. (2009). Cystic fibrosis. *The Lancet 373* (9678), 1891–1904.

[21] Reyes, A., Wu, M., et al. (2013). Gnotobiotic mouse model of phage–bacterial host dynamics in the human gut. *Proc. Natl. Acad. Sci. 110* (50), 20236–20241.

[22] Rice, P., Longden, I., et al. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet. 16* (6), 276–277.

[23] Römling, U., Kader, A., et al. (2005). Worldwide distribution of *Pseudomonas aeruginosa* clone C strains in the aquatic environment and cystic fibrosis patients. *Environ. Microbiol. 7* (7), 1029–1038.

[24] Römling, U., Wingender, J., et al. (1994). A major *Pseudomonas aeruginosa* clone common to patients and aquatic habitats. *Appl. Environ. Microbiol. 60* (6), 1734–1738.

[25] Schmieder, R. and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One 6* (3), e17288.

[26] Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics 27* (6), 863–864.

[27] Skennerton, C.T., Imelfort, M., et al. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res. 41* (10), e105.

[28] Smoot, M.E., Ono, K., et al. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics 27* (3), 431–432.

[29] Speert, D.P., Campbell, M.E., et al. (2002). Epidemiology of *Pseudomonas aeruginosa* in cystic fibrosis in British Columbia, Canada. *Am. J. Respir. Crit. Care Med. 166* (7), 988–993.

[30] Truong, D.T., Franzosa, E.A., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods 12* (10), 902–903.

[31] Willner, D., Haynes, M.R., et al. (2012). Spatial distribution of microbial communities in the cystic fibrosis lung. *ISME J. 6* (2), 471–474.

[32] Winstanley, C., Langille, M.G.I., et al. (2009). Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res. 19* (1), 12–23.

CHAPTER 5

CRISPR Surveillance of a Panmictic Global Population of *Pseudomonas aeruginosa* Phage via a Novel Prototyping Method

Abstract

The demographics of the human virome remain elusive to the high-throughput molecular examination that is the mainstay of microbial ecology. Here we develop the sequence-specific CRISPR adaptive immune system as a surveillance tool to track virus demographics within and between microbial populations associated with the opportunistic human pathogen *P. aeruginosa*. We find that CRISPR spacers match the majority of known sequenced *P. aeruginosa* phages, as well as predicted prophage in well-characterized strains. We use these spacers to identify specific sequence tags that can be used to recognize and track phage populations in a high-resolution, high-throughput way. We use the distribution of these sequence tags to show that *P. aeruginosa* and its phages are globally distributed and highly mobile between the human and environmental spheres. The abundance of sequence signatures matching particular phage groups demonstrates the abundance of interactions between this group and the diversity of *P. aeruginosa* strains that inhabit all environments. Application of this large library of spacers as a surveillance tool to track phage elements can be used beyond CRISPR typing to follow phage dynamics independent of their hosts with important implications for tracking virulence islands and antibiotic resistance, as well as making personalized predictions for phage therapy in this potent human pathogen.

Introduction

Viral infection is known to have considerable impact on the evolution of microbial communities in all environments including the human microbiome [3,42,48,49,50]. Among the many human body sites where the microbial community can impact health is the cystic fibrosis lung. Cystic fibrosis is among the most common life-shortening human genetic disorders, and while the disease is genetic in origin, the vast majority of morbidity and mortality stems from persistent microbial infections in the lung [19]. Among the most widespread and damaging of these pathogens is *Pseudomonas aeruginosa*, which plays a key role in cystic fibrosis morbidity and mortality [28]. *P.*

aeruginosa is a ubiquitous opportunistic pathogen, and cystic fibrosis patients are believed to primarily be colonized by strains encountered in their environment, which subsequently adapt to the lung environment and thrive [28]. Eradicating an established *P. aeruginosa* infection via antibiotic therapy can be difficult to impossible, leading to consideration of phage therapy as an alternative treatment method. Numerous studies have shown the therapeutic promise in vitro and in vivo in animal models [46]. For phages to become a safe and effective treatment strategy, understanding their potential interactions with *P. aeruginosa* is a critical step.

Comparisons of even small numbers of *P. aeruginosa* genomes has revealed a dynamic variable genome replete with horizontally transferred elements, many of which are prophages and phage-like elements [35,41]. This variable genome can occupy 10% of the otherwise highly conserved *P. aeruginosa* genome and contains genes which this pathogen can use to maintain its presence in the human body, including virulence factors and antibiotic resistance genes [37]. Phages represent a major pathway for the transfer of these critical genes; for example, the temperate cytotoxin-converting phage phiCTX encodes a toxin shown to increase *P. aeruginosa* virulence in a mouse model [4]. Other prophages have affected varying functions important for colonization and persistence, including cell adhesion, resistance to phagocytosis, and exopolysaccharide digestion for biofilm remodeling [37]. Notably, prophage also play an important role in the Liverpool epidemic strains (LES), which are responsible for 10% of cystic fibrosis-associated infections in the United Kingdom [39]. These strains are adept at colonizing the lung, display increased antibiotic resistance, and are associated with worse clinical outcomes [25]. Some of the colonization advantages of the LES strains have been shown to lie in integrated prophage in the LES genome. These elements contain genes homologous to phages D3, D3112, F10, and Pf1, and disrupting three of these prophage has been show to put the strain at a competitive disadvantage relative to its wild-type ancestor in a rat lung chronic infection model [61]. Some of these integrated phages have also been shown to retain their lytic activity and may affect *P. aeruginosa* density in chronic cystic fibrosis lung infections [30].

Tracking elements such as phages, which can significantly impact longevity, virulence, and persistence of their bacterial hosts, has great impacts on understanding human disease dynamics. However, because phages are highly modular, diverse and represent a relatively small portion of the DNA in any environment, investigating their ecology using molecular tools is difficult. Luckily, *P. aeruginosa* employs the CRISPR-cas immune system as its own mechanism for surveillance of mobile elements. The CRISPR system consists of arrays of alternating palindromic repeats interspersed with short DNA fragments called spacers and a series of CRISPR-associated (cas) genes. CRISPR spacers frequently match segments (protospacers) of mobile genetic elements such as viruses, plasmids, and transposons [43]. Upon encountering a foreign genetic element, one or more new spacers matching that element can be added to one end of the repeat-spacer array, known as the leader [5]. The complete repeat-spacer array is transcribed and processed into crRNAs containing individual spacers, which are used to guide ribonucleoprotein complexes to corresponding protospacers when the matched element is seen again. These complexes inactivate and/or degrade the targeted DNA, preventing infection or transfer [10]. In response to this threat, viruses can evade CRISPR targeting by acquiring random mutations in their protospacers [21,56].

CRISPR systems are divided into three types and numerous subtypes based on their cas gene content [38]. *P. aeruginosa* is known to harbor two subtypes of the Type I CRISPR system in its genome: I-E and I-F. Type I-F CRISPRs are considerably more common than Type I-E, appearing in 33% of genomes versus 3% for Type I-E in a study of 122 clinical isolates [14]. The Type I-F system has been shown to be fully functional as an immune system, conferring immunity to multiple temperate phages and adding new spacers in response to challenge with a lytic phage [13]. In addition to these genomically encoded CRISPRs, a Type I-C locus has recently been identified on an integrative and conjugative element present in some *P. aeruginosa* strains [8]. Spacers from all of these CRISPR subtypes have been shown to match known phage, with approximately a quarter of unique spacers uncovered in multiple studies matching phage or prophage [8,14], indicating that *P. aeruginosa* has recorded many encounters with phage in its CRISPR arrays.

Given the evidence for phage and prophage activity, phage influence on *P. aeruginosa* in the cystic fibrosis lung microbiome, and the potential for *P. aeruginosa* CRISPRs to record a history of phage interactions, we investigated the extent of interaction with phage by assessing CRISPR spacer content in a large pool of *P. aeruginosa* genomes from varied environments. We found a diverse array of spacers matching a broad variety of sequenced phages and leveraged these spacers to investigate the population structure of these phages on a global scale. Overall targeting of phages by host was investigated with respect to viral lifestyle, host environment, and host geographic location, revealing interactions between host and virus.

Results

A total of 1,184 *P. aeruginosa* genome sequences were compiled for analysis. Nearly half of the strains in the dataset were isolated from cystic fibrosis patients (560), with the remaining strains originating from human samples not related to cystic fibrosis (421), human samples of undetermined cystic fibrosis status (43), environmental samples (34), strains derived from laboratory experiments (14), animal samples (1), or unknown origins (111). These strains were collected in 26 countries and isolated over a range of 25 years (Figure 5.1). The complete seven-locus multi-locus sequence type was determined for 1,058 *P. aeruginosa* strains (Figure 5.2), showing that cystic fibrosis patients and other sources are intermingled on the tree, supporting established findings that cystic fibrosis patients are colonized by a variety of environmentally derived *P. aeruginosa* strains [34]. Strains isolated from the same country, even from the same clinic, are also widely dispersed across the tree. Exceptions to this include known epidemic strains of *P. aeruginosa* [31], which are clustered by geographic location (Figure 5.2). To compare the phylogeny of CRISPR spacers to core genome phylogeny, a maximum parsimony tree of spacer presence was constructed for all CRISPR-containing strains (Figure 5.3). CRISPR spacers were organized into 729 unique repeat-spacer array loci, present in 754 of the 1,184 strains. We find that CRISPRs resolve differences among strains that are identical by MLST, but the CRISPR tree remains consistent with the MLST tree. While some individual spacers are shared between strains of different sequence types, we do not observe the same entire array in strains of different sequence types. This indicates that

although they evolve relatively rapidly, arrays do not move between lineages through recombination on this time scale.

Protospacer typing of P. aeruginosa *phage*

To determine whether spacers could be used to classify phage, we assembled a group of 92 sequenced phage of *P. aeruginosa*, as well as 88 predicted prophage regions from 13 fully sequenced *P. aeruginosa* genomes. This set includes diverse prophage, with a total sequence length 10.7 Mbp and a pan-genome size of over 7.2 Mbp. We divided our known and predicted phage library into phage clusters based on the fraction of the phage genomes aligned in pairwise BLAST searches (see methods). We identified 31 phage clusters with at least two members, along with 15 clusters with a single representative (Figure 5.4A).

Many of the larger clusters correspond well to established *P. aeruginosa* phage groups [29,54]. Cluster03 and Cluster08 contain the D3112-like and B3-like temperate transposable phages, respectively, while Cluster07 includes D3 and several related prophages. The phiKMV-like podophages occupy Cluster06, and two species of lytic myophages form Cluster05 and Cluster10. Interestingly, the filamentous phages are split, with Pf1 and several prophages forming Cluster04, while Pf3 did not cluster with any other phages in this dataset. Cluster09 contains phage F10, as well as eight prophage, including related prophage found in LES strains; F10-like phage genes have also been identified in the variable genome of ST111 *P. aeruginosa*, a widespread nosocomial sequence type [58]. Many other known phages form smaller clusters, often including related prophages (see Table A.1 for a complete list of cluster members). Eleven of the 31 clusters, including the two containing the most members, contained only predicted prophage.

In total we identified 3,152 unique spacer sequences from 1,184 *P. aeruginosa* strains. A rarefaction curve of spacer sequences reveals that despite a broad sampling of available *P. aeruginosa* sequence data, not all spacers in the population have been observed (Figure 5.5A), showing this species holds a highly diverse collection of spacers. To assess size of the phage genome space sampled by spacers targeting the same phage

region, we quantified how often the same protospacer region was sampled by CRISPR spacers (100% identity over at least 20 bp). When combining spacers with this similarity, we retain 2,949 of the initial 3,152 spacer sequences. A rarefaction curve built from these independent acquisitions shows that the same phage region is rarely targeted repeatedly by these hosts (Figure 5.5B).

Many of the spacers identified match *P. aeruginosa* phage or prophage, with 1,592 matching sequenced phage, 1,042 matching predicted prophage and 846 matching both. A total of 1364 spacers match neither phage nor prophage in this dataset, indicating that CRISPRs are sampling a set of foreign genetic elements much larger than the set of known *P. aeruginosa* phages. All but 24 (18 independently sequenced, 6 predicted) of these 180 phage sequences contained a protospacer matched by at least one of the identified spacers.

We next explored the feasibility of using protospacer matched to infer phage identities by analyzing the relationship between overall phage genome similarity and similarity of the protospacer content of the genomes. We determined a protospacer type, or the set of all spacer-matched sequences present in each phage genome, and then we clustered phage by similarity of protospacer sets. For the most part, members of the same phage genome cluster were also found in the same protospacer cluster, with some exceptions. In one case, while most phages from genome cluster 02 grouped together by protospacers in protospacer cluster 05, one phage was pulled into protospacer cluster 06 with phages from genome clusters 25 and 39, as it shared more protospacers with those phages than with those in its genome cluster (Figure 5.4B). In two instances, a subset of members of a genome cluster with an especially high proportion of shared protospacers was pulled into an independent prototype cluster (genome cluster 01 into protospacers clusters 04 and 15; genome cluster 06 into protospacer clusters 07 and 19; see Figure 5.4B). While most genome clusters contain only one protospacer type, the majority of protospacer clusters (10/19) contained members of multiple genome clusters (Figure 5.4B), indicating that some protospacers are common to distantly related phages. Given this propensity for dispersed protospacers, we wanted to know how broadly these protospacers were conserved, and if there were also protospacers which could differentiate taxonomically

related groups of phages. For each genome cluster, the protospacers present were classified into one of five groups (Figure 5.4C): cluster-specific (found in all phages of that genome cluster and only that genome cluster), cross-cluster (found in all phages of the genome cluster, plus in at least one phage of another genome cluster), distributed (found in some phages of the genome cluster and also in some phages of other genome clusters), intermediate (found only in one genome cluster, but not in all phages), and individual (found only in a single phage in the entire dataset). In 21 out of 27 protospacer-containing phage genome clusters, 50% or more of protospacers were found to be present in multiple genome clusters (cross-cluster or distributed; Figure 5.4D), which led us to hypothesize that more conserved regions might be more highly targeted by spacers; however, we found that protospacers were not present in more strains in this dataset than non-matched regions of the phage genome (Figure 5.6). While 64% of spacers with matches match only phages in a single cluster, a small number of spacers appear to broadly target up to 10 genome clusters, with five spacers targeting nine or more (Figure 5.7). Interestingly, these five protospacers all match the same phage region, annotated as a conserved hypothetical protein.

In addition to these inter-cluster protospacers, all but one protospacer-containing phage cluster has at least one spacer exclusive to that cluster, and over half the phage genome clusters had at least one cluster-defining protospacer (Figure 5.4D, Table A.2). However, protospacers are not present more frequently in more conserved regions of the genome (Figure 5.6). Indeed, for all but three clusters, more spacers match variable regions not present in all strains than core regions conserved cluster-wide (Figure 5.8).

Spatial and temporal distribution of phage

Finally, we used the distribution of CRISPR spacers within our *P. aeruginosa* isolate database to survey the phage distribution in time and space. We began by constructing a maximum parsimony tree based on the presence and absence of CRISPR spacers and mapping on the location and environment in which they were found (Figure 5.3). It is clear from this tree that unique spacer content does not differentiate strains from different countries; however, it is possible that spacer content could be more similar between strains from the same environment and geographic location, with strains living closer

together or in the same environment harboring more similar spacers. To investigate this, we examined the proportion of spacers shared between each pair of strains within and between environments and found very similar levels of shared spacers within as well as between environments (Figure 5.9). Shared spacers were also compared to the geographic distance between collection sites. We found only a very weak negative correlation (Pearson's $r = -0.06$, $p < 2.2 \times 10^{-16}$) between distance and shared spacers (Figure 5.10), showing that there is no appreciable relationship between environment or geography and spacer presence in *P. aeruginosa*.

In addition to shared spacers between strains, the frequency of CRISPR presence was also analyzed with respect to the isolation environment and country (Figure 5.11). Strains isolated from cystic fibrosis patients were significantly more likely to have CRISPRs compared to strains from other environments (Chi-squared = 18.36, df = 1, p-value = 1.8×10^{-5}) and lab-derived strains were more likely to lack CRISPRs (Chi-squared = 18.86, df = 1, p-value = 1.4×10^{-5}; chi-square for each environment vs. all other environments with Bonferroni correction). No country of origin was significantly associated with increased or decreased CRISPR presence.

To determine if some phage genome clusters were more frequently targeted by the CRISPR system, we quantified the number of protospacers per base pair of sequence in each phage genome. Some phage groups are targeted significantly more or less frequently than most other groups (Figure 5.12); most notably clusters 03 and 18, which have significantly more protospacers per base pair than 26 and 25 out of 30 other clusters, respectively (one-way ANOVA with Games-Howell test). In total, six clusters were more highly targeted than two-thirds of the groups identified, revealing that some phage types are indeed more heavily sampled by *P. aeruginosa* CRISPR systems, while others have very few protospacers. In fact, four groups lack any protospacers matched by spacers in our library. Notably, the most highly-targeted groups consist of predominantly temperate phages (Figure 5.4A). Across the entire set of phages, temperate phages contain significantly more protospacers than phages characterized as lytic (Figure 5.13, Student's t test, $p < 2 \times 10^{-9}$).

To learn if this differential targeting of phage groups was related to their environment, we compared the number of hosts with a spacer matching each group to the number of hosts from each environment matching that group. Only five comparisons proved to be significantly different ($p < 0.05$), and in all cases, the group had fewer protospacers matched by spacers from that environment than the total population (Student's t test with Bonferroni correction; Table 5.1). We also considered that hosts in proximity to one another may see similar phage communities and thus have spacers targeting the same phages. We clustered hosts by geographic distance and found that most phage genome clusters appear in multiple host geographic clusters, with a consistently high mean distance between geographic clusters each phage genome cluster is found in (Figure 5.14). Genome clusters matching relatively geographically close hosts have very few protospacers, indicating that they are rarely targeted by CRISPRs, and thus have less potential to be matched by hosts from different locations. Among phage groups with moderate to high numbers of protospacers, no group appears restricted to a smaller geographic area, showing little association between geographic distance and acquisition of immunity to the same phage species and indicating that phage groups are widely distributed globally.

While the majority of phage clusters are distributed across distance and environment, we hypothesized that hosts may encounter these phage clusters at different times relative to other clusters, as phages move among environments or geographic regions. To quantify how relatively recently hosts had encountered particular spacers or spacers associated with each phage cluster, we assigned each spacer a newness value equal to its proximity to the leader end of its repeat-spacer array. With respect to environment, no individual spacer was found to be more or less recent in any environment; however, spacers matching four phage clusters did show significantly different levels of newness (Figure 5.15). In three of the clusters, the same trends appear: matching spacers appear most recently in environmental strains, less recently in strains from cystic fibrosis patients, and least recently in non-cystic-fibrosis clinical isolates. In the fourth cluster, spacers are found more recently in non-cystic-fibrosis strains than in those from cystic fibrosis environments (one-way ANOVA with Games-Howell test).

Geographically, there is a weak positive correlation between distance between hosts and absolute difference in array position of the same spacer (Pearson's r=0.16, p < 2.2e-16), indicating little association between geographic distance and relative time of phage encounter on a spacer-by-spacer basis. When spacers are partitioned by the phage groups they match, most correlations vary from weak negative to weak positive (Pearson's r -0.33 to 0.24); however, in Cluster13 there is a significant moderate positive correlation between spacer position distance and geographic distance (r=0.55) (Table 5.2). These limited correlations indicate that hosts in closer proximity generally do not necessarily encounter and acquire immunity to phages at similar times, although the moderate correlation present in Cluster13 indicates that it may be an exception to this rule Additionally, we analyzed the leadermost spacer in each array, which represents the phage the host most recently acquired immunity to, for hosts in each distance-based cluster. All geographic clusters contained hosts with leadermost spacers matching multiple phage clusters, with no geographic area clearly favoring a particular phage cluster (Figure 5.16).

Discussion

We have investigated the extent of interactions between *P. aeruginosa* and its phages through analysis of CRISPR spacer and protospacer content in a large pool of strains isolated from a diverse range of environments. This analysis has demonstrated high spacer diversity in *P. aeruginosa* strains across environments, with identical CRISPR arrays rare between strains and little evidence of correlation between environment or location and CRISPR spacer similarity. We also found that the global *P. aeruginosa* spacer pool contains spacers matching the majority of sequenced *P. aeruginosa* phages, as well as predicted prophages from complete *P. aeruginosa* genomes, making the spacer library a practical tool for investigating phage population structure. Many spacers targeted sequence that appeared in multiple phage clusters, though nearly all phage clusters contained protospacers which were unique to that cluster and half had at least one cluster-specific spacer which appears in all strains, making it a marker for host encounter with that cluster. Some phage clusters were found to be much more frequently targeted than others, particularly those containing predominantly temperate phage. This targeting

was not associated with location and was only linked to environment in a small number of cases. Additionally, we found that hosts from different locations generally do not encounter related phages at the same relative time, and phage clusters showed limited evidence of immune acquisition at different times in different environments.

Evidence has long indicated that cystic fibrosis patients are colonized by *P. aeruginosa* strains from their environment which subsequently adapt to the lung niche [28]. Our work supports this view, with an MLST phylogeny placing strains from cystic fibrosis patients alongside those isolated from other human-associated and non-human-associated environments rather than in distinct environmentally defined clades. Certain epidemic strains are exceptions to this rule. LES-like strains are found exclusively in cystic fibrosis patients from the UK and Canada. ST111 and ST235, a pair of multidrug resistant MLST types which are commonly found in hospital-acquired infections [22,59,62] were predominantly present in non-cystic-fibrosis clinical samples and absent from environmental samples, though ST235 appears in two cystic fibrosis patients from Copenhagen. The other set of strains which appears to drive the formation of environment-specific clades is a longitudinal dataset of cystic fibrosis patients from Marvig *et al*. [40]; however, this is primarily due to clones at the MLST and spacer levels which are isolated from the same patient over time. A phylogeny built on spacer content also indicated that related strains inhabit multiple environments, with nearly half of well-supported clades containing strains from multiple environments. Similarly, no clear geographic pattern emerges from either the MLST or spacer tree, where presence in multiple countries was more common than presence in a single nation for well-supported clades. This lack of a geographic or environmental pattern in spacer presence shows that *P. aeruginosa* strains in varied environments around the world see similar phages, indicating that, like their hosts, they are broadly distributed globally and move freely between human-associated and non-human-associated environments.

The widespread high spacer diversity observed in this dataset highlights the importance of the CRISPR system in *P. aeruginosa*-phage interactions. Similar to a recent study focused on clinical isolates of *P. aeruginosa* [8], we found a diverse array of CRISPR spacer content which was consistent with the MLST phylogeny of the strains. Unlike

Belkum *et al*., we focused our spacer extraction solely on chromosomally encoded CRISPR loci of Types I-F and I-E, excluding the Type I-C CRISPR loci infrequently encoded in mobile elements. While we find potential Type I-C arrays in 23 of our 1,187 genomes, we find our spacer library thoroughly samples *P. aeruginosa* phage space without these arrays. The identified spacers cover protospacers present in the majority of sequenced *P. aeruginosa* phages as well as predicted prophages in the most complete genomes of *P. aeruginosa* available to date, indicating that the CRISPR system provides immunity to a broad spectrum of phages. Shared protospacer sequences between phage clusters, even those with low overall genome similarity, are common and reflect the modular nature of some *P. aeruginosa* phages; for example, of the six well-studied prophages of the epidemic strain LESB58, two pairs of prophages share regions 7.5 kb and 13.5 kb in length [61]. Targeting these shared regions could be advantageous for a host strain, allowing acquisition of immunity to phages it has not encountered previously.

In some instances, we observed significantly higher or lower than anticipated CRISPR presence. In strains from cystic fibrosis patients, CRISPRs were present more frequently than in strains from other environments, pointing to a potentially important role in antiviral defense for *P. aeruginosa* strains inhabiting the cystic fibrosis lung. Interestingly, strains from the same environment do not share more spacers with one another than they do with strains from other environments, and spacer similarity did not strongly correlate with proximity of hosts to one another. This further indicates that *P. aeruginosa* phages are broadly distributed spatially and among habitats, such that all types of hosts encounter all types of phage in a global panmictic population, or that the potential protospacer sequence space is large enough that hosts rarely sample the same part of the phage genome. We found evidence supporting both of these hypotheses; protospacers which were exclusively found in single phage clusters were also distributed widely across different environments and locations, which supports the idea that phage are ubiquitously distributed, and there are few instances of closely related spacers appearing in different array contexts (Figure 5.5B), showing that hosts rarely gain immunity to a phage in the same way.

Among phage clusters, some were found to be targeted by hundreds of spacers, while four were targeted by none at all. By far the most highly targeted phage group we found was Cluster03, which contained 259 unique protospacers; the individual phages in this cluster have approximately five protospacers per kb of genome. This cluster contains D3112 and related temperate transposable phages, whose mosaic genomes have been heavily shaped by horizontal gene transfer among Mu-like and lambda-like phages [29]. Their overrepresentation in CRISPR loci may relate to their entry mechanism; these phage use Type IV pili as their receptors [12,51], and these pili are important for motility on solid surfaces and in viscous environments, and play a crucial role in biofilm structure [44]. In culture, virus-resistant *P. aeruginosa* mutants have been shown to typically delete the pilus to prevent viral attachment; however, in resistant strains where the pilus is not deleted, CRISPR spacers are added to confer immunity [13]. Selective pressure to maintain motility and proper biofilm structure in environments such as the cystic fibrosis lung may have pushed CRISPR immunity over pilus-compromising resistance mechanisms, leading to heavy targeting of these phages. Consistent with this hypothesis, other phages which are known to use Type IV pili for entry, including B3-like phage (Cluster08), and F116 (Cluster25) also have moderate to high numbers of protospacers (Figure 5.12).

Outside of Cluster03, temperate phages were more heavily targeted than lytic phages, as previously observed in a smaller set of *P. aeruginosa* spacers, where all 132 spacers which perfectly matched phage matched temperate phage or prophage [14]. This skewed targeting could indicate that CRISPR immunity is used less frequently for defense against lytic phage, with other resistance mechanisms being used instead. It is also possible that integrated phages are used competitively between *P. aeruginosa* strains, such as observed in experimental populations [11,33], and CRISPRs provide a way to specifically target phages used in this way by competitor strains. However, these highly targeted clusters were highly targeted in all environments, indicating that these competitive dynamics are not restricted to any one host habitat.

Despite the broad distribution of hosts across the globe and in different environments, we saw no strong correlation between geographic distance and how recently hosts had

acquired immunity to phage groups. Hosts showed more recent acquisition in certain environments of spacers to four phage groups; in three cases, environmental strains had added spacers more recently than cystic fibrosis or clinical strains not linked to cystic fibrosis. Two of these clusters, Cluster03 and Cluster08, are comprised of D3112-like and B3-like phages, while Cluster02 contains only prophage. The fourth cluster shows a reversal of this pattern, with non-cystic-fibrosis strains adding spacers more recently than cystic fibrosis strains; however, only four spacers match this cluster overall, and this limited targeting may contribute to this result more than a meaningful difference in timing of spacer acquisition. While the position of a spacer in its array can only tell us how recently a host gained immunity to a phage cluster relative to its encounters with other phages, this pattern indicates that *P. aeruginosa* strains in human-associated environments have encountered other types of phages since they last gained fresh immunity to these three phage types, while in the environment they are among the more recently encountered phages.

Using variable CRISPR content for fine-scale tracking of infectious microbes has been proposed for clinical isolates of *P. aeruginosa* [8], as well as in other bacteria [1,17,24,27]; here, we propose to instead apply the wealth of information encoded in CRISPR arrays to prototyping, a method which can differentiate, track, and untangle the population structure of the phages rather than their hosts. Prototyping allows us to leverage sequencing data to track which phage demographics in dynamic microbial populations such as *P. aeruginosa*, and could be applied to any microbial system where CRISPRs adequately sample the phage population. Protospacers found exclusively in a single phage or phage cluster hold promise for identifying viral variants in the host genome; these cluster- or phage-specific spacers could be used to quickly classify a newly identified element. Prototyping can be used to identify and track integrated phage or other mobile elements linked to critical determinants of pathogenic success, such as virulence and antibiotic resistance islands, or any variable segments of the host genome which distinguish stains from one another, such as the F10-like phage genes and mobile elements present in the widespread clinical *P. aeruginosa* isolate type ST111 [8,58] and the advantageous prophages of LES epidemic cystic fibrosis strains[61]. Prototyping also holds exciting implications for personalized medicine by enhancing the efficacy of phage

therapy. Spacer libraries matching candidate phages for phage therapy could be used to identify hosts harboring immunity to the phage, thus predicting its effectiveness at eliminating infection. Such a strategy could be used on a broad level, to identify classes of phage which are generally infrequently targeted by hosts, or on an individual level to provide a patient with maximally effective phage to which the infecting strain lacks immunity.

These results depict a global population of *P. aeruginosa* and phage where many phage types are continuously circulating across a broad geographic area and in multiple environments. Host repeat-spacer arrays bear evidence of encounters with many types of phage without a clear spatial or environmental pattern. However, the significantly higher targeting of particular phage groups, largely consisting of temperate phages, reveals that hosts have different immune responses to different phage types. Using a large library of spacers extracted from a large dataset spread across time, space and sample type allowed us to see how these phage were differentially targeted on a global scale rather than examining individual host-phage interactions. Applying this type of surveillance to other host-virus systems could similarly reveal novel patterns in CRISPR targeting and viral population structure.

Methods

Host dataset selection

The set of *P. aeruginosa* genomes and CRISPR arrays analyzed in this paper includes data from the following sources. Reads associated with 458 *P. aeruginosa* strains cultured from patient samples collected from the Copenhagen Cystic Fibrosis Center at the University Hospital, Rigshospitalet, Denmark [40] were retrieved from the NCBI Sequence Read Archive (Accession ERP004853) and assembled as described below. Assembled genomes and sequencing reads of 24 *P. aeruginosa* strains described in Dettman *et al.* [20] were kindly provided by the authors. Assembled genomes of 388 strains described in [36] were obtained from GenBank (BioProject accession PRJNA264310). All other complete and draft-stage *P. aeruginosa* genomes contained in the NCBI Nucleotide database were retrieved in September 2014 (310 genomes). In

addition to complete genomes, CRISPR arrays from [14] were downloaded from NCBI (45 sequences). Three additional sets of CRISPR arrays were taken from metagenomic sequence of three cystic fibrosis patient sputum samples kindly provided by Katrine Whiteson and Yan Wei Lim of San Diego State University. For all strains, basic metadata, including isolation location, sampling date, environment, and epidemic strain status were collected where possible.

Quality filtering and genome assembly

For all samples with sequencing reads available, reads were trimmed and quality filtered using Prinseq 0.20.4 [53]. Reads were trimmed from the left and right ends using a 5 nt sliding window with a minimum quality score of 30. Reads were retained if they had a mean quality score of 30 and 1% or fewer ambiguous bases. The minimum read length was set to approximately two-thirds the anticipated read length, or 66 nt. Draft assemblies were generated with MIRA 4.0 [16] using the genome, denovo, and accurate parameters.

Multi-locus sequence typing

An established panel of seven markers [18] was used for MLST analysis of all strains. MLST loci were identified by BLASTn [2] of a representative known allele obtained from the *Pseudomonas aeruginosa* pubMLST website (http://pubmlst.org/paeruginosa/) [32] against genomes or contigs. The best BLAST hit for each MLST locus was then BLASTed against a database of all known alleles for that locus, also obtained from the *Pseudomonas aeruginosa* MLST website. Exact matches to one of these known alleles were assigned that allele's ID number; hits with lower identity or incomplete coverage of the locus were investigated manually, and any identified as novel alleles were assigned new ID numbers >10,000. Strains with inconclusive MLST alleles were removed from further MLST analysis. A maximum-likelihood tree of concatenated MLST markers was constructed with RAxML [55] using the rapid bootstrapping algorithm plus maximum likelihood and GTRGAMMA nucleotide substitution model with 100 bootstrap replicates.

CRISPR identification and spacer extraction

CRISPR loci were identified in genomes or contigs via BLASTn of known *P. aeruginosa* CRISPR repeats [14]. Parameters were adjusted for the short search sequence and to maximize hits covering the entire repeat length as follows: "-word _size 7 -gapopen 3 -gapextend 2 -reward 1 -penalty -1". The minimum percent identity was set to 80 to allow for detection of degenerate repeat sequences. Additionally, hits shorter than 24 bp in length were filtered from the results. Sequences with a repeat of the same type both up- and downstream in the same orientation and that were less than 40 bp away from any neighboring spacers were considered spacers and extracted. Each new unique spacer identified was numbered sequentially. A spacer rarefaction curve was computed in QIIME [15].

In addition to spacer extraction, CRISPR repeat-spacer array ranges were declared as all consecutive repeats and spacers in the same orientation less than 500 bp away from one another. Groups of repeats and spacers on different contigs, on the same contig/genome in different orientations, or on the same contig/genome but separated by more than 500 bp were considered separate arrays.

Clone correction

To avoid biasing the dataset with multiple identical or nearly identical strains from longitudinal sampling in the Marvig *et al*. data, strains isolated from the same patient with identical MLST loci and CRISPR spacer content were condensed to a single representative strain. This resulted in the removal of 365 strains from the dataset.

Maximum parsimony spacer analysis

A presence/absence matrix of all 3,152 spacers in all CRISPR-containing strains was constructed and used to generate a maximum parsimony tree using PAUP* 4.0b10 [57]. One hundred bootstrap replicates were performed. For analysis of metadata within clades, well-supported clades were defined as the largest clades with bootstrap support of 50% or greater. Well-supported clades consisting solely of identical strains from the

Marvig *et al*. dataset were removed from analysis as noted in the clone correction section above, as they would have been reduced to a single strain.

Phage dataset selection and protospacer identification

Genomes of all phage identified as infecting *P. aeruginosa* were downloaded from the NCBI Nucleotide database on June 23, 2015, totaling 92 unique phages. Predicted prophages were identified in 13 complete *P. aeruginosa* genomes using VirSorter [52]. Results were kindly provided by the authors prior to public availability of the program. Predictions from all three confidence categories were used, and 88 potential prophages were identified. All phages were classified according to lifestyle (lytic, temperate, or non-lytic) based on their descriptions in the literature. Together, these 180 phage and prophage were used for all phage-related analyses.

Determination of phage pan-genome size

The size of the *P. aeruginosa* phage pan-genome was determined using a custom python script. This script hierarchically creates a non-redundant pseudo-genome based on a set of input DNA sequences. From this original collection of DNA sequences, the largest sequence is used as a scaffold for the non-redundant genome, and BLAST is performed to identify components of the remaining pool that are shared with the scaffold. Shared segments are removed from the pool, the largest contiguous sequence remaining in the pool is appended to the scaffold, and this cycle is repeated until the pool of sequences is empty.

Protospacers in phage genomes were identified via BLASTn of all spacer sequences against the sequence of interest. The parameter "-task blastn-short" was used due to short query length. A minimum e-value of 0.01 was used to capture incomplete and imperfect matches, allowing up to four mismatches over a full-length match. Partial-length matches were extended to cover the full spacer length using clDB [7].

Assignment of phages to genome clusters

To assign phages to clusters, all phage genomes were compared to one another using BLASTn (e=0.001). For each pair of genomes, the proportional length alignment (PLA),

or total length aligned by BLAST over the length of the query, was calculated and used as our measure of similarity between phages. MCL [23] was used to cluster phage groups into networks with edges weighted by PLA, with a minimum PLA cutoff of 0.2.

Clustering hosts by distance

For each host where the sampling location was known, coordinates were determined using the GeoNames geographical database [26]. If the city or other specific location was not provided, the capital or seat of the smallest known political region was used instead. Hosts were divided into 33 geographical clusters using the modularity calculation in Gephi [6], which implements the Louvain Method for finding subcommunities within larger networks [9]. Edges were weighted by the inverse distance between each host pair.

Statistical analysis

All statistical tests were performed in R versions 3.2.2-3.2.4 [47]. Games-Howell tests were performed using the userfriendlyscience package [45]. Plots were generated in R using the ggplot2 package [60].

Figures and Tables

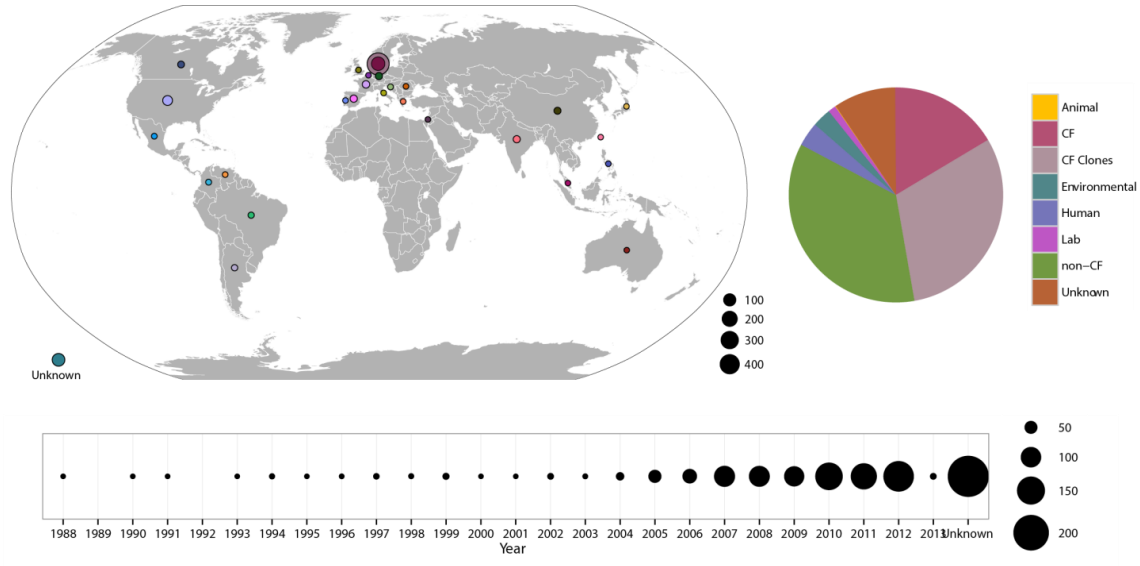


Figure 5.1. Isolation environment, geographical location, and year of isolation for 1,184 *P. aeruginosa* sequences used in this study. "Human" refers to strains known to have been isolated from a human where the cystic fibrosis status of the person is unknown. "CF clones" denotes strains identical at the MLST and CRISPR level, which were removed from further analysis.

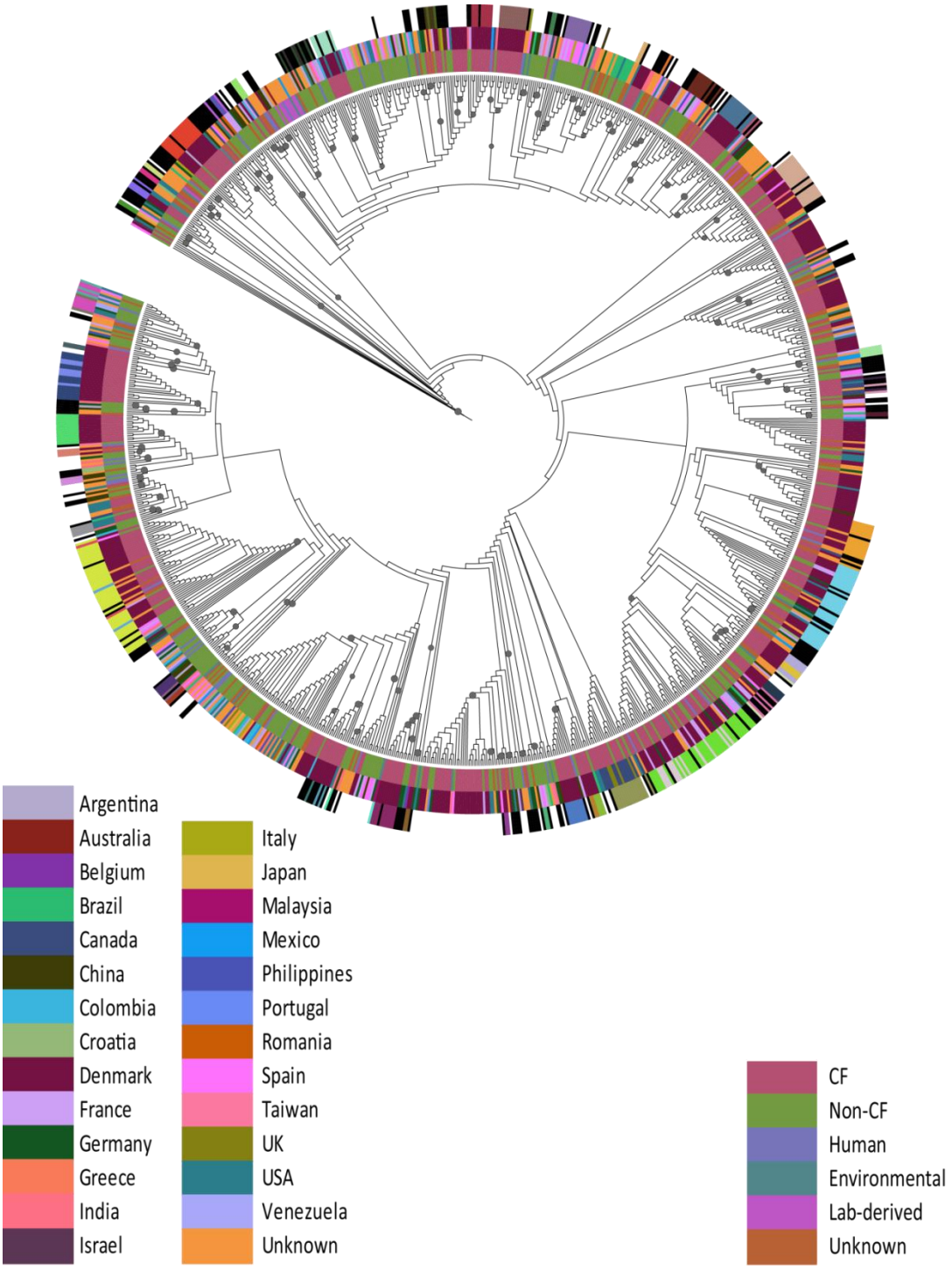


Figure 5.2. Concatenated maximum likelihood MLST tree of *P. aeruginosa* strains. Bootstrap values >70 are indicated by circles. Inner ring: environment; middle ring: country of isolation, outer ring: CRISPR-type. Strains with no CRISPR spacers are shown in white; strains with a unique CRISPR-type are shown in black.

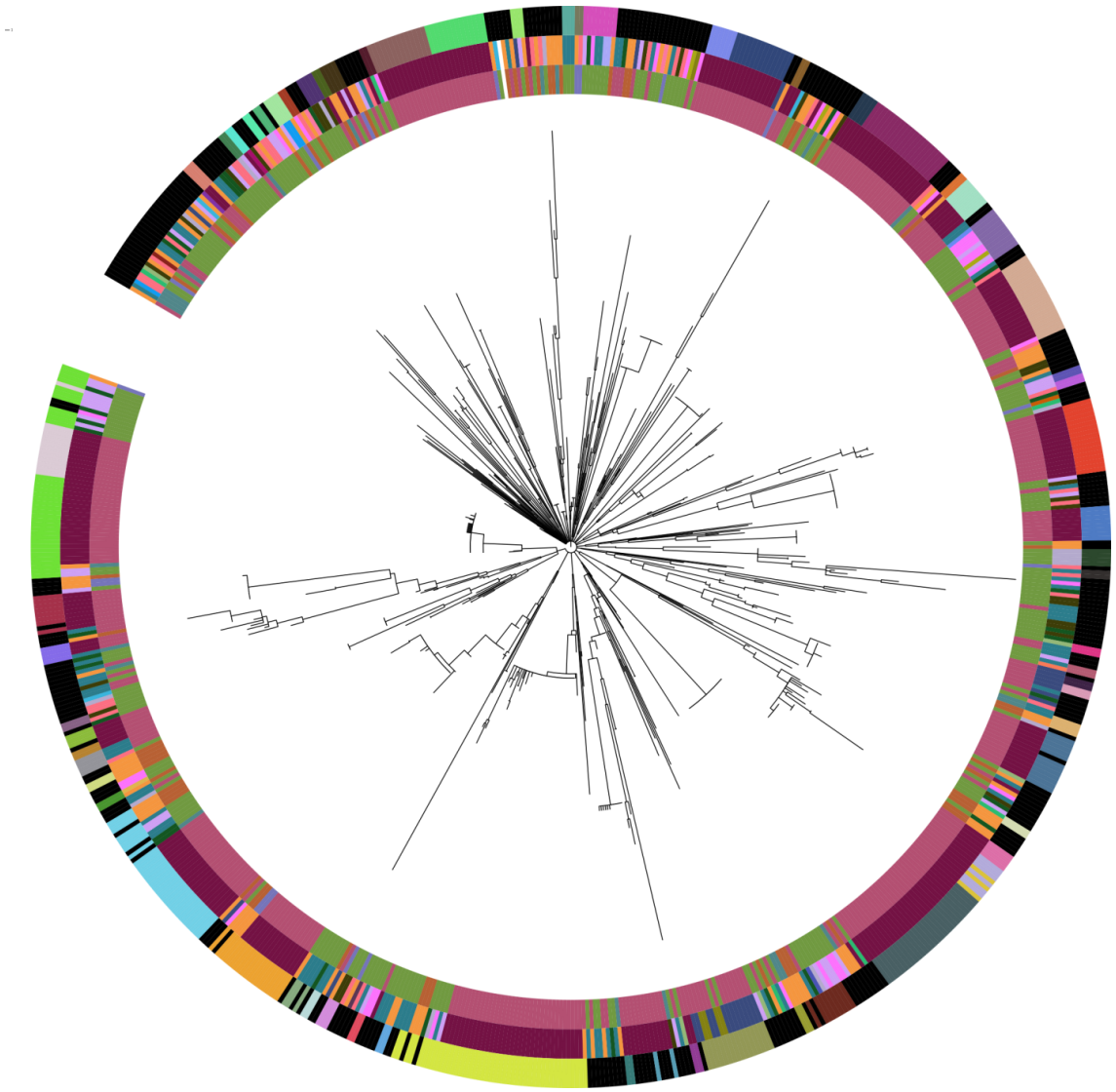


Figure 5.3. Maximum parsimony tree of spacer content in CRISPR-containing strains. Inner ring: environment; middle ring: country of isolation, outer ring: CRISPR-type. Strains with a unique CRISPR-type are shown in black.

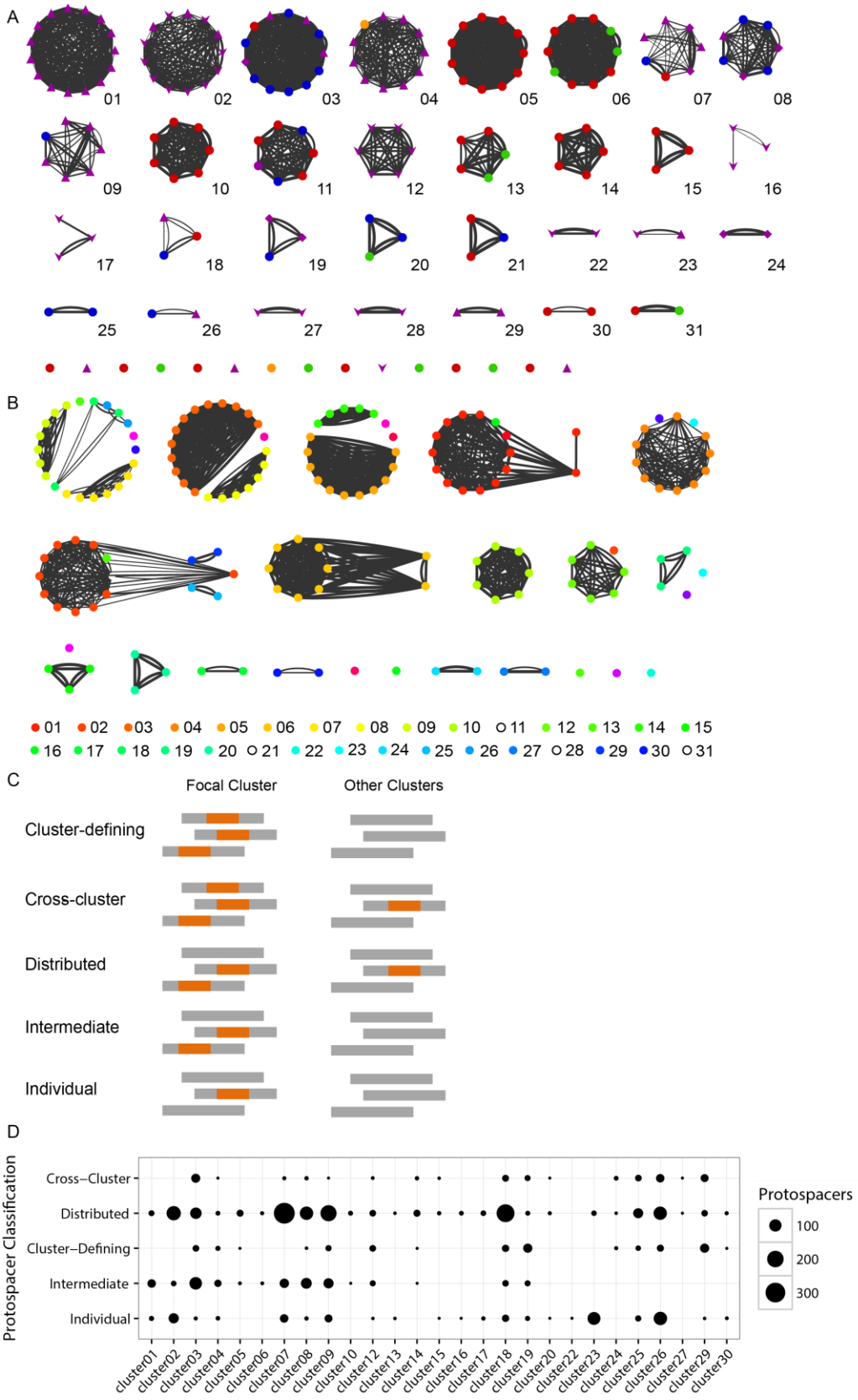


Figure 5.4. Phage clusters and defining protospacers.

Figure 5.4. (cont.)

A. Identified phage genome clusters. Numerical identifiers are to the lower right of each cluster. Each node represents a phage sequence. Line width represents sequence shared between phages; color indicates phage lifestyle (red, lytic; blue, temperate; purple, prophage; orange, non-lytic; green, uncertain). B. Protospacer type clusters. Members of the same genome cluster share a node color and are connected by lines, as in A, showing genome clusters that are pulled together into the same protospacer cluster, or pulled apart into separate protospacer clusters. The colors representing each genome cluster number are shown in the legend at the bottom. C. Classifications of protospacer types based on presence within and between phage genome clusters. Gray bars represent individual phages; orange segments represent protospacers. Protospacers in the cluster-defining, intermediate, and individual classes are useful for discriminating between genome clusters. D. Classification of protospacers matching each phage cluster. Dots show number of protospacers in each phage cluster at each level of conservation.

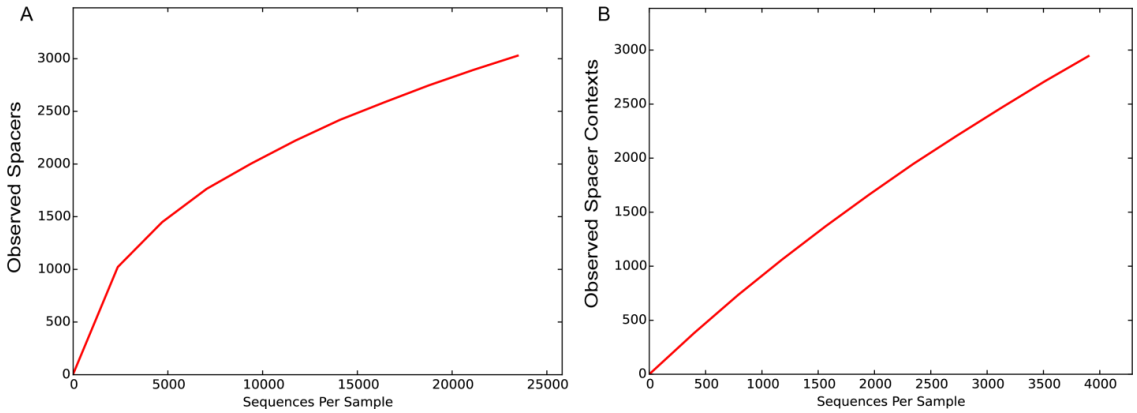


Figure 5.5. Spacer rarefaction curves. A. Rarefaction curve of spacers identified in 1,184 *Pseudomonas aeruginosa* genomes. B. Rarefaction curve of spacers observed in unique contexts (different trailer-side spacer).

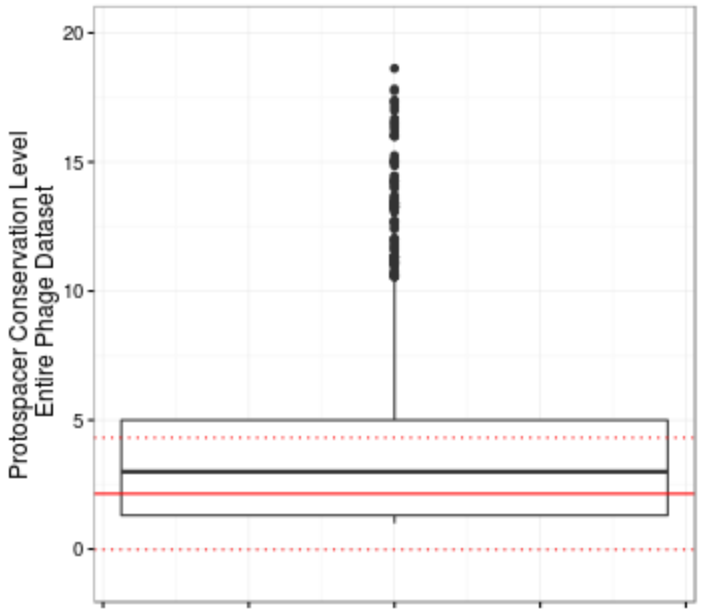


Figure 5.6. Protospacers are not more conserved than general phage sequence. The black boxplot represents the number of phage genomes containing each protospacer, with the mean number of genomes containing non-protospacer phage sequence +/- sd is shown in red. The center line of the boxplot represents the median; the upper and lower lines mark the first and third quartiles, respectively. Whiskers extend to 1.5x the interquartile range, points outside this range are shown as black dots.

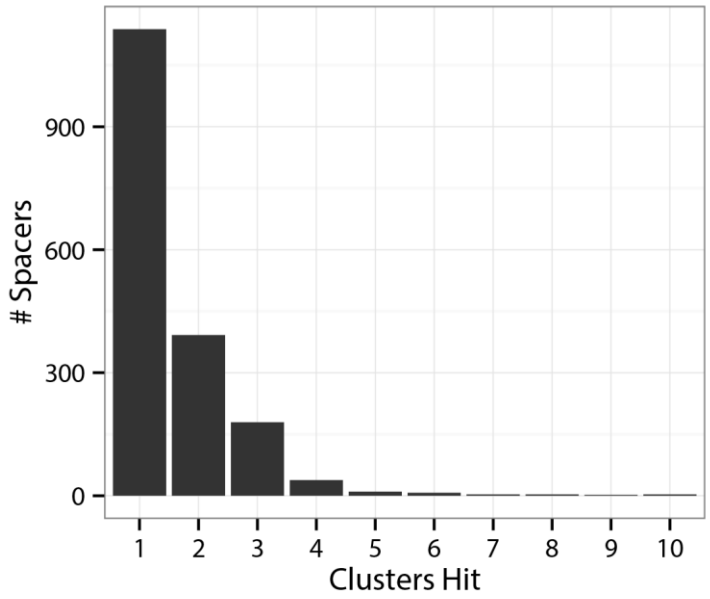


Figure 5.7. Distribution of the number of phage genome clusters hit by spacers. Only phage genome clusters with two or more members are considered.

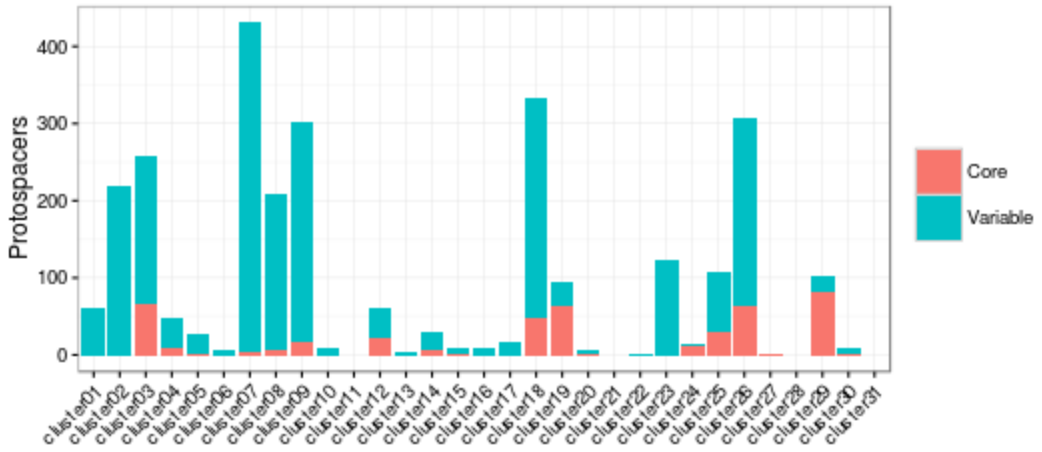


Figure 5.8. Spacers targeting core or variable sequence in each phage cluster. For each cluster, core (red) is defined as protospacer sequence present in all phages in that cluster; variable (blue) sequence is missing from at least one cluster member.

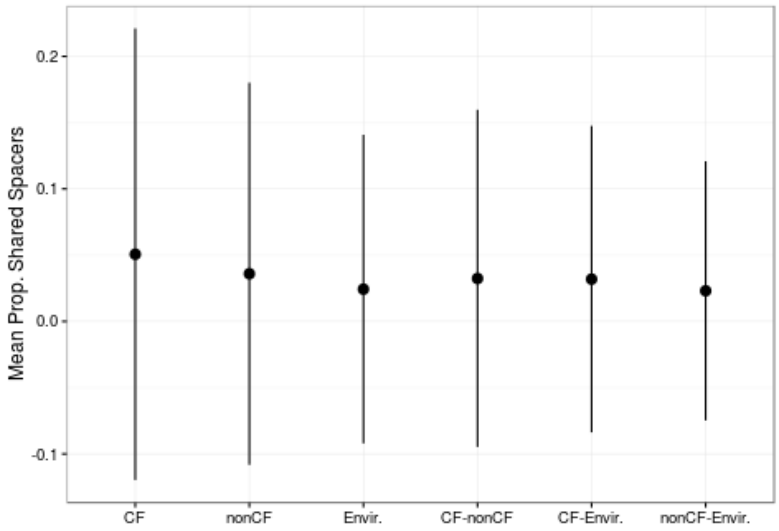


Figure 5.9. Proportion of spacers shared between *P. aeruginosa* host pairs vs. the environment from which they were isolated. Error bars show +/- sd.

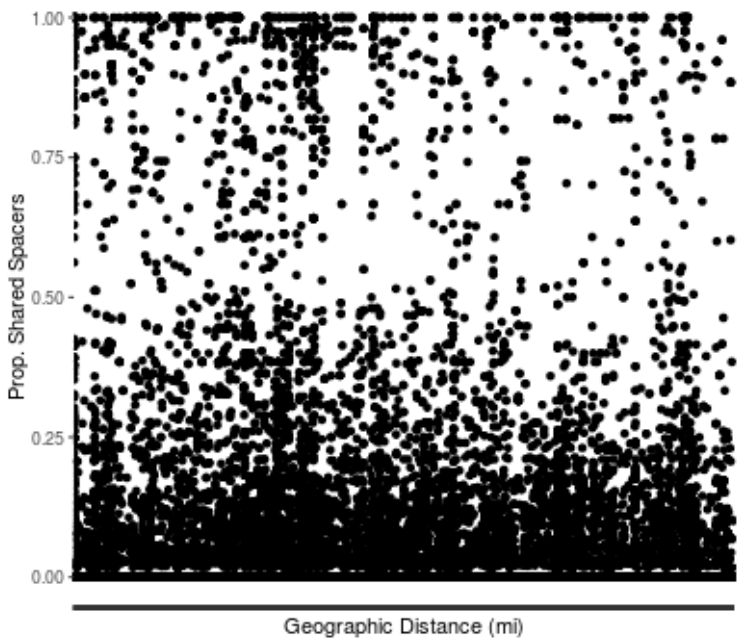


Figure 5.10. Proportion of spacers shared between *P. aeruginosa* host pairs vs. the geographic distance between isolation locations. Only a very weak correlation (Pearson's r= -0.06) is observed.

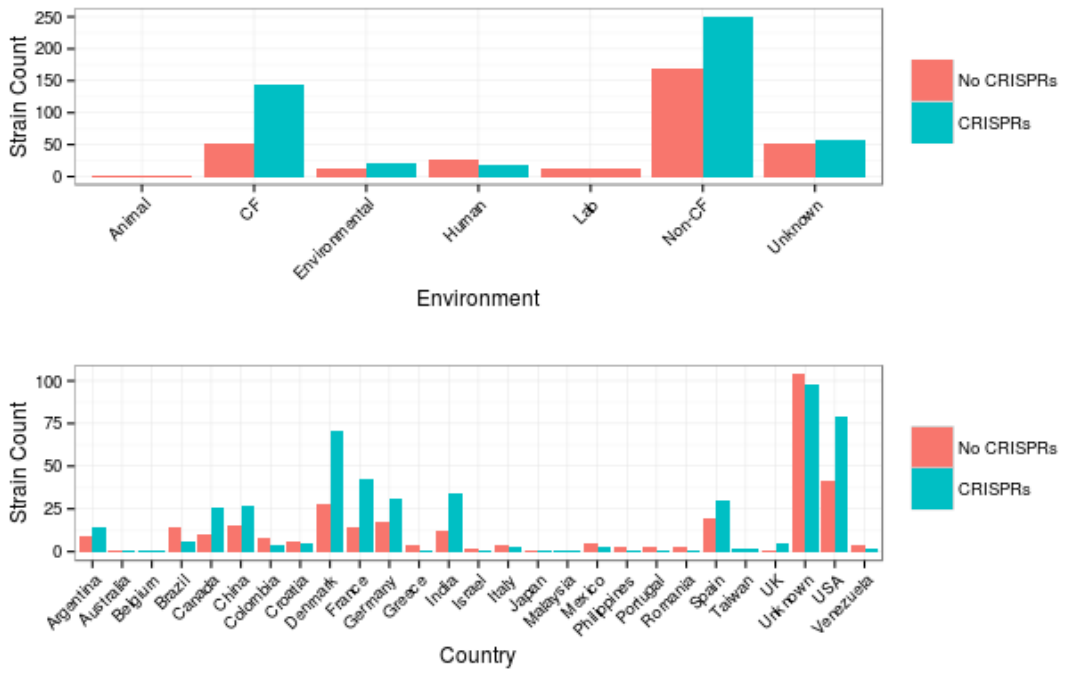


Figure 5.11. CRISPR-containing versus CRISPRless strains in each environment or country of isolation. CRISPR-containing strains were significantly more common in strains from cystic fibrosis ($p=1.82x10^{-5}$), while CRISPRless strains were significantly more common in lab-derived strains ($p=1.41x10^{-5}$). (Chi-square for each environment or country vs. all others with Bonferroni correction).

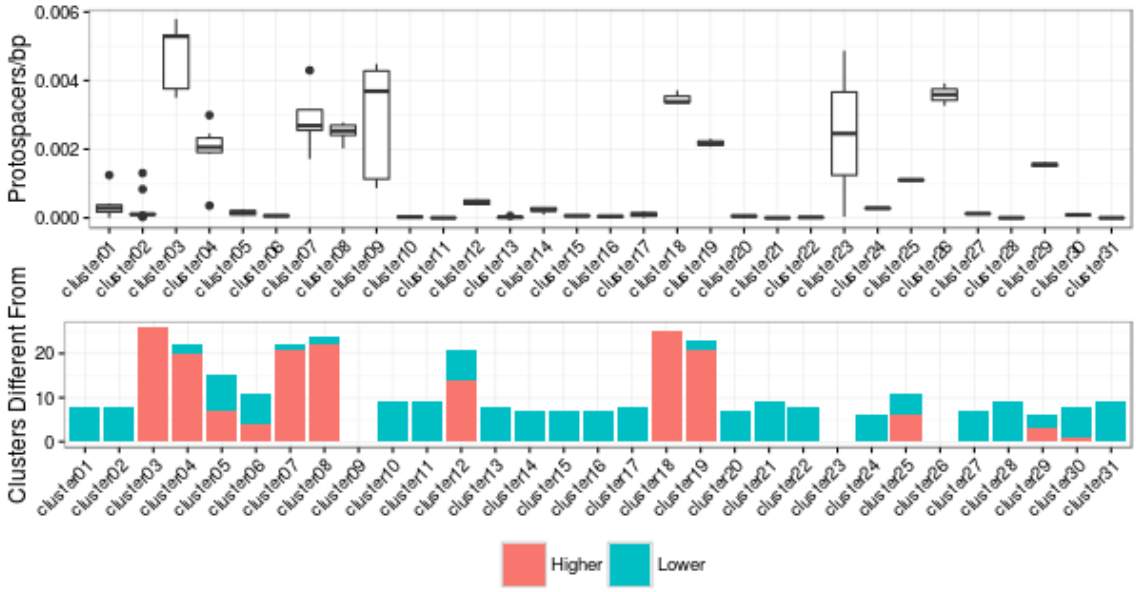


Figure 5.12. CRISPR targeting of phage clusters. Upper: Protospacers per base pair of sequence for each phage cluster. The center line of the box represents the median; the upper and lower lines mark the first and third quartiles, respectively. Whiskers extend to 1.5x the interquartile range, points outside this range are shown as black dots. Lower: Count of other clusters from which each phage cluster has a significantly higher (red) or lower (blue) number of protospacers per base pair (one-way ANOVA with Games-Howell test, $p < 0.05$).

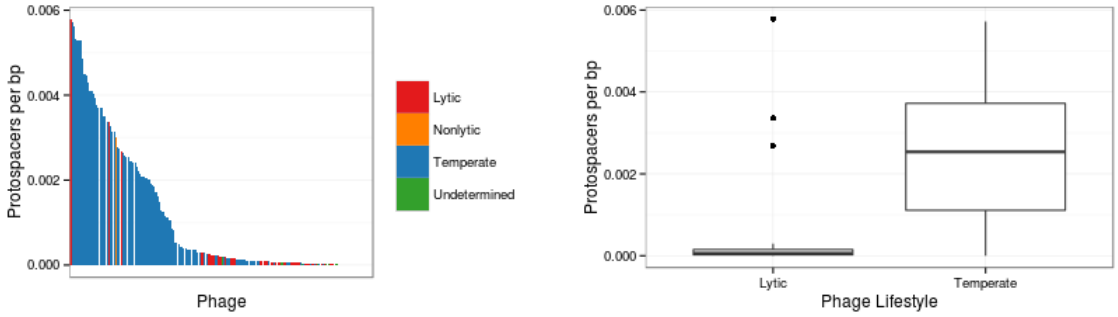


Figure 5.13. Spacers target temperate phages. Left: Number of protospacers per base pair of sequence in 180 sequenced and predicted phage categorized by lifestyle. Right: Protospacers per base pair is significantly higher in temperate phages than lytic phages (Student's t-test, $p=1.56x10^{-9}$). The center line of the box represents the median; the upper and lower lines mark the first and third quartiles, respectively. Whiskers extend to 1.5x the interquartile range, points outside this range are shown as black dots.

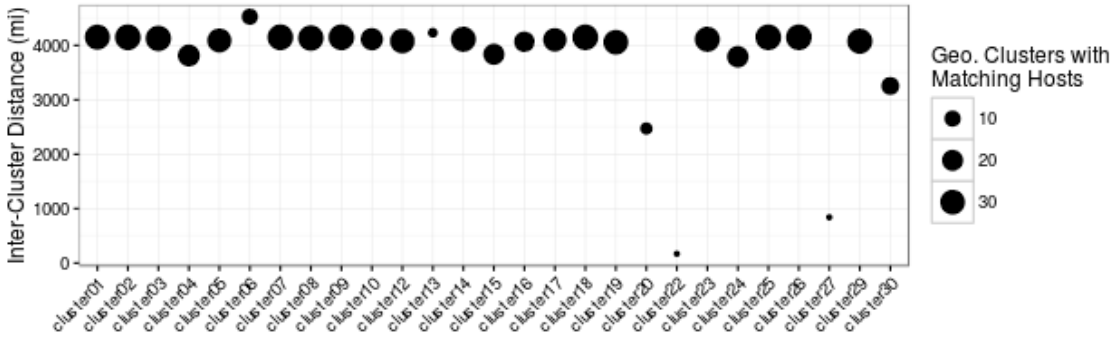


Figure 5.14. Mean distance between host geographic clusters matched by phage genome clusters is generally high. Genome clusters are shown on the x-axis, with the mean distance between geographic clusters which contain hosts matching each phage genome cluster on the y-axis. Dot size is scaled to the number of geographic clusters with hosts matching each phage cluster.

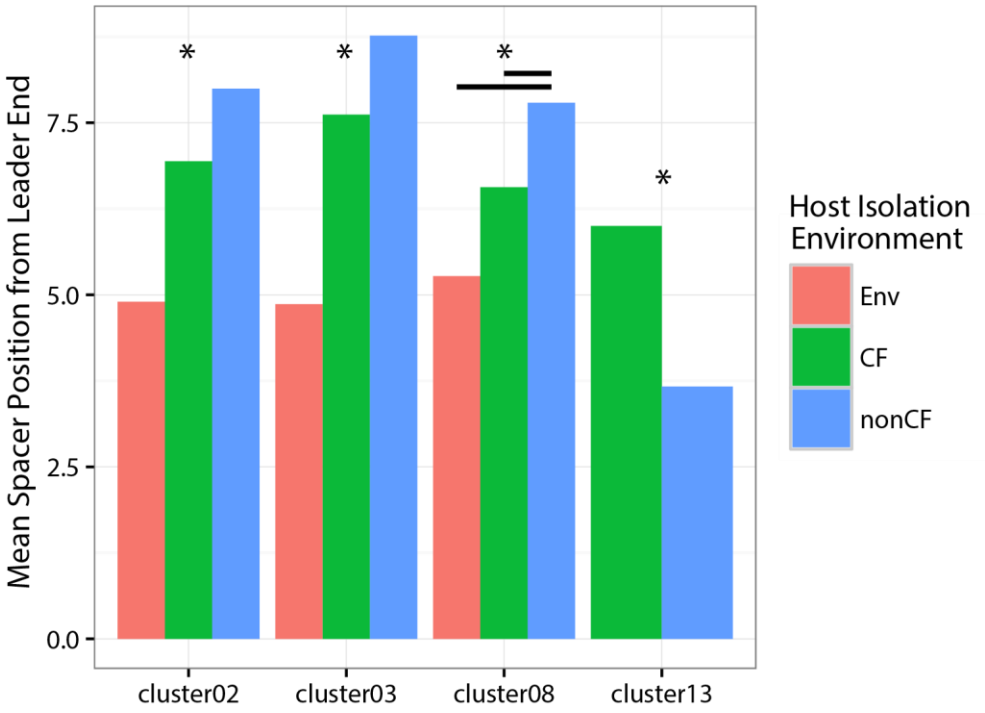


Figure 5.15. Spacer distance from leader varies across environments for four phage groups. For Clusters 02, 03, and 13, all environments are significantly different from one another; in Cluster08, horizontal bars designate which environments are significantly different. *, p <0.01.

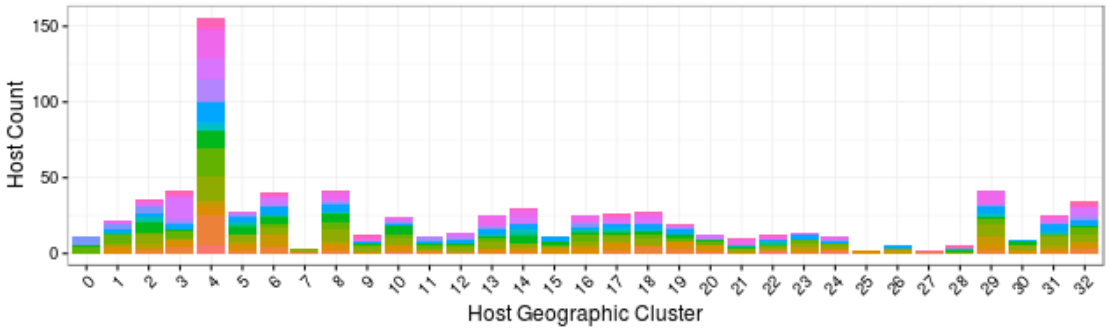


Figure 5.16. Host geographic clusters target multiple phage clusters with their recently-acquired spacers. Each color corresponds to a phage cluster; the height of each colored band corresponds to the number of hosts with CRISPR spacers that match that phage cluster.

Table 5.1. Phage clusters matched by significantly fewer spacers from an environment than the total phage population.

Cluster	Environment	p	Mean spacers per host (Cluster)	Mean spacers per host (All)
cluster04	nonCF	2.59E-05	1.35	2.51
cluster06	Environmental	1.59E-69	0	0.18
cluster13	Environmental	1.59E-69	0	0.18
cluster20	Environmental	1.59E-69	0	0.18
cluster27	Environmental	1.59E-69	0	0.18

Table 5.2. Correlation between host geographic distance and position in repeat-spacer array between spacers matching phage clusters.

Cluster	r	p
cluster01	-0.07	3.82E-21
cluster02	0.02	7.49E-38
cluster03	0.05	0
cluster04	0.25	5.28E-71
cluster05	0.05	1.18E-06
cluster06	0.05	6.92E-01
cluster07	0.02	2.94E-66
cluster08	0.09	0
cluster09	0.01	2.95E-34
cluster10	0.10	1.21E-07
cluster12	0.02	1.02E-09
cluster13	0.55	4.37E-04
cluster14	0.03	7.18E-04
cluster15	0.11	3.93E-08
cluster16	0.01	7.85E-01
cluster17	0.10	3.71E-08
cluster18	0.01	3.85E-33
cluster19	0.02	1.47E-06
cluster20	0.15	3.93E-01
cluster22	NA	NA
cluster23	0.02	2.02E-17
cluster24	0.10	3.04E-05
cluster25	0.03	3.21E-24
cluster26	0.01	1.11E-27
cluster27	-0.33	5.19E-01
cluster29	0.03	2.27E-08
cluster30	-0.11	2.48E-01

References

[1] Abadia, E., Zhang, J., et al. (2010). Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. 10* (7), 1066–1074.

[2] Altschul, S.F., Gish, W., et al. (1990). Basic local alignment search tool. *J. Mol. Biol. 215* (3), 403–410.

[3] Andersson, A.F. and Banfield, J.F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science 320* (5879), 1047–1050.

[4] Baltch, A.L., Smith, R.P., et al. (1994). *Pseudomonas aeruginosa* cytotoxin as a pathogenicity factor in a systemic infection of leukopenic mice. *Toxicon 32* (1), 27–34.

[5] Barrangou, R., Fremaux, C., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science 315* (5819), 1709–1712.

[6] Bastian, M., Heymann, S., et al. (2009). Gephi: An open source software for exploring and manipulating networks. *Third International AAAI Conference on Weblogs and Social Media*, (2009).

[7] Bautista Chavarriaga, M.A. (2016). Viral diversity and host-virus interactions in the model crenarchaeon *Sulfolobus islandicus*. Ph.D dissertaton, University of Illinois at Urbana-Champaign.

[8] Belkum, A. van, Soriaga, L.B., et al. (2015). Phylogenetic distribution of CRISPR-cas systems in antibiotic-resistant *Pseudomonas aeruginosa*. *mBio 6* (6), e01796-15.

[9] Blondel, V.D., Guillaume, J., et al. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* (2008), P10008.

[10] Brouns, S.J.J., Jore, M.M., et al. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science 321* (5891), 960–964.

[11] Brown, S.P., Le Chat, L., et al. (2006). Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. *Curr. Biol. CB 16* (20), 2048–2052.

[12] Budzik, J.M., Rosche, W.A., et al. (2004). Isolation and characterization of a generalized transducing phage for *Pseudomonas aeruginosa* strains PAO1 and PA14. *J. Bacteriol. 186* (10), 3270–3273.

[13] Cady, K.C., Bondy-Denomy, J., et al. (2012). The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *J. Bacteriol. 194* (21), 5728–5738.

[14] Cady, K.C., White, A.S., et al. (2011). Prevalence, conservation and functional analysis of *Yersinia* and *Escherichia* CRISPR regions in clinical *Pseudomonas aeruginosa* isolates. *Microbiology 157* (2), 430–437.

[15] Caporaso, J.G., Kuczynski, J., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods 7* (5), 335–336.

[16] Chevreux, B., Wetter, T., et al. (1999). Genome sequence assembly using trace signals and additional sequence information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics*, (1999), 45–56.

[17] Comas, I., Homolka, S., et al. (2009). Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One 4* (11), e7815.

[18] Curran, B., Jonas, D., et al. (2004). Development of a multilocus sequence typing scheme for the opportunistic pathogen *Pseudomonas aeruginosa*. *J. Clin. Microbiol. 42* (12), 5644–5649.

[19] Cystic Fibrosis Foundation. (2015). *Patient Registry Annual Data Report 2014*.

[20] Dettman, J.R., Rodrigue, N., et al. (2013). Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. 110* (52), 21065–21070.

[21] Deveau, H., Barrangou, R., et al. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol. 190* (4), 1390–1400.

[22] Edelstein, M.V., Skleenova, E.N., et al. (2013). Spread of extensively resistant VIM-2-positive ST235 *Pseudomonas aeruginosa* in Belarus, Kazakhstan, and Russia: a longitudinal epidemiological and clinical study. *Lancet Infect. Dis. 13* (10), 867–876.

[23] Enright, A.J., Dongen, S.V., et al. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res. 30* (7), 1575–1584.

[24] Fabre, L., Zhang, J., et al. (2012). CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS One 7* (5), e36995.

[25] Fothergill, J.L., Walshaw, M.J., et al. (2012). Transmissible strains of *Pseudomonas aeruginosa* in cystic fibrosis lung infections. *Eur. Respir. J. 40* (1), 227–238.

[26] GeoNames. (2016). *GeoNames*. Unxos GmbH, Switzerland.

[27] Ginevra, C., Jacotin, N., et al. (2012). *Legionella pneumophila* sequence type 1/Paris pulsotype subtyping by spoligotyping. *J. Clin. Microbiol. 50* (3), 696–701.

[28] Hauser, A.R., Jain, M., et al. (2011). Clinical significance of microbial infection and adaptation in cystic fibrosis. *Clin. Microbiol. Rev. 24* (1), 29–70.

[29] Hertveldt, K. and Lavigne, R. (2008). Bacteriophages of *Pseudomonas*. In B.H.A. Rehm, ed., *Pseudomonas*. Wiley-VCH Verlag GmbH & Co. KGaA, 255–291.

[30] James, C.E., Davies, E.V., et al. (2015). Lytic activity by temperate phages of *Pseudomonas aeruginosa* in long-term cystic fibrosis chronic lung infections. *ISME J. 9* (6), 1391–1398.

[31] Jeukens, J., Boyle, B., et al. (2014). Comparative genomics of isolates of a *Pseudomonas aeruginosa* epidemic strain associated with chronic lung infections of cystic fibrosis patients. *PLoS One 9* (2), e87611.

[32] Jolley, K.A. and Maiden, M.C. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics 11* (1), 595.

[33] Joo, J., Gunny, M., et al. (2006). Bacteriophage-mediated competition in *Bordetella* bacteria. *Proc. R. Soc. B Biol. Sci. 273* (1595), 1843–1848.

[34] Kidd, T.J., Ritchie, S.R., et al. (2012). *Pseudomonas aeruginosa* exhibits frequent recombination, but only a limited association between genotype and ecological setting. *PLoS One 7* (9), e44199.

[35] Klockgether, J., Cramer, N., et al. (2011). *Pseudomonas aeruginosa* genomic structure and diversity. *Front. Microbiol. 2*, 150.

[36] Kos, V.N., Déraspe, M., et al. (2015). The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrob. Agents Chemother. 59* (1), 427–436.

[37] Kung, V.L., Ozer, E.A., et al. (2010). The accessory genome of *Pseudomonas aeruginosa*. *Microbiol. Mol. Biol. Rev. 74* (4), 621–641.

[38] Makarova, K.S., Wolf, Y.I., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol. 13* (11), 722–736.

[39] Martin, K., Baddal, B., et al. (2013). Clusters of genetically similar isolates of *Pseudomonas aeruginosa* from multiple hospitals in the UK. *J. Med. Microbiol. 62* (Pt_7), 988–1000.

[40] Marvig, R.L., Sommer, L.M., et al. (2015). Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat. Genet. 47* (1), 57–64.

[41] Mathee, K., Narasimhan, G., et al. (2008). Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proc. Natl. Acad. Sci. 105* (8), 3100–3105.

[42] Modi, S.R., Lee, H.H., et al. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature 499* (7457), 219–222.

[43] Mojica, F.J.M., Díez-Villaseñor, C., et al. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol. 60* (2), 174–182.

[44] O'Toole, G.A. and Kolter, R. (1998). Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. *Mol. Microbiol. 30* (2), 295–304.

[45] Peters, G. (2016). *userfriendlyscience: Quantitative analysis made accessible*. http://userfriendlyscience.com/.

[46] Pires, D.P., Boas, D.V., et al. (2015). Phage therapy: a step forward in the treatment of *Pseudomonas aeruginosa* infections. *J. Virol. 89* (15), 7449–7456.

[47] R Core Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

[48] Reyes, A., Wu, M., et al. (2013). Gnotobiotic mouse model of phage–bacterial host dynamics in the human gut. *Proc. Natl. Acad. Sci. 110* (50), 20236–20241.

[49] Rodriguez-Brito, B., Li, L., et al. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J. 4* (6), 739–751.

[50] Rohwer, F. and Thurber, R.V. (2009). Viruses manipulate the marine environment. *Nature 459* (7244), 207–212.

[51] Roncero, C., Darzins, A., et al. (1990). *Pseudomonas aeruginosa* transposable bacteriophages D3112 and B3 require pili and surface growth for adsorption. *J. Bacteriol. 172* (4), 1899–1904.

[52] Roux, S., Enault, F., et al. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ 3* , e985.

[53] Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics 27* (6), 863–864.

[54] Sepúlveda-Robles, O., Kameyama, L., et al. (2012). High diversity and novel species of *Pseudomonas aeruginosa* bacteriophages. *Appl. Environ. Microbiol. 78* (12), 4510–4515.

[55] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics 30* (9), 1312–1313.

[56] Sun, C.L., Barrangou, R., et al. (2013). Phage mutations in response to CRISPR diversification in a bacterial population. *Environ. Microbiol. 15* (2), 463–470.

[57] Swofford, D.L. (2003). *Phylogenetic Analysis Using Parsimony (*and Other Methods).* Sinauer Associates, Sunderland, Massachusetts.

[58] Turton, J.F., Wright, L., et al. (2015). High-resolution analysis by whole-genome sequencing of an international lineage (sequence type 111) of *Pseudomonas aeruginosa* associated with metallo-carbapenemases in the United Kingdom. *J. Clin. Microbiol. 53* (8), 2622–2631.

[59] Viedma, E., Villa, J., et al. (2014). Draft genome sequence of colistin-only-susceptible *Pseudomonas aeruginosa* strain ST235, a hypervirulent high-risk clone in Spain. *Genome Announc. 2* (5), e01097-14.

[60] Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

[61] Winstanley, C., Langille, M.G.I., et al. (2009). Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res. 19* (1), 12–23.

[62] Witney, A.A., Gould, K.A., et al. (2014). Genome sequencing and characterization of an extensively drug-resistant sequence type 111 serotype O12 hospital outbreak strain of *Pseudomonas aeruginosa*. *Clin. Microbiol. Infect. 20* (10), O609–O618.

CHAPTER 6

Concluding Remarks and Future Directions

CRISPR Diversity in Simulated Populations and Translation to Natural Populations

Through the use of an eco-evolutionary model of host-virus coevolution mediated by CRISPR immunity, we were able to peer into the evolution of CRISPR diversity and how it can lead to complex host populations with distributed CRISPR immunity to the resident viruses. In these simulated populations, phenotypically identical strains with underlying genotypic diversity contribute to a stable host population which is better able to limit or eliminate growth of predatory viruses. These findings provide evolutionary context for observations of high CRISPR diversity in species such as *Sulfolobus islandicus* [2] and crucially provides novel metrics for assessing distributed immunity in experimental and natural populations.

While the findings and metrics derived from these simulations are valuable, they require further testing on real-world microbe-virus communities to determine if the observed CRISPR diversity also manifests in living systems, and if host stability and viral instability follow distributed immunity as the model predicts. To this end, we chose two discrete human-associated microbial communities: the vaginal microbiomes of pregnant women and the lung microbiomes of cystic fibrosis patients. In both, we found surprisingly limited diversity among CRISPR-containing strains within an individual human host, which stands in stark contrast to the simulated populations and previous observations in *S. islandicus* populations. This low diversity may result from many aspects of the host-viral ecology in these environments. They may encounter fewer viruses overall, or fewer lytic viruses, which would decrease evolutionary pressure to maintain diversified CRISPR immunity. The viruses present in these systems may also be predominantly temperate, with the potential to carry valuable genes for virulence, antibiotic resistance, or many other advantageous processes. Many non-CRISPR mechanisms of viral resistance exist in microbes [5], and these particular environments may favor those mechanisms over CRISPR immunity.

CRISPRs in the Vaginal Microbiome

To better understand the low CRISPR diversity in these environments, further work is required to characterize the host and viral populations, their CRISPR-mediated interactions, and other viral resistance mechanisms. In the vaginal microbiome, sequencing and analysis of paired bacterial and viral metagenome samples would be valuable for following the trajectory of the viral population in parallel to tracking CRISPR changes in the hosts. Conducting sampling outside of pregnancy may also help clarify the presence of CRISPR diversity. The vaginal microbiome changes markedly during pregnancy [1,7,8], and accordingly we observed major shifts in both species and CRISPR-type abundance over the course of pregnancy and the postpartum period in the subjects in our study. These shifts may have led to loss of diversity due to bottlenecking of CRISPR-containing species which were reduced in abundance; samples taken during a time period when the community is less subject to disturbance may better assess CRISPR diversity in the vaginal microbiome under less disruptive circumstances.

CRISPRs in the Cystic Fibrosis Lung Microbiome

While our initial foray into CRISPR diversity in the cystic fibrosis lung microbiome uncovered no within-sample diversity in *Pseudomonas aeruginosa* CRISPRs, evidence of lytic action by temperate phages in LES-infected cystic fibrosis patients [4] as well as continued research into phage therapy as a treatment for *P. aeruginosa* [3] underscore the importance of understanding host-virus interactions in this environment. CRISPRs in particular may be an important mechanism of resistance to phages which use Type IV pili as part of their entry mechanisms. These pili are involved in biofilm formation, which is an important part of *P. aeruginosa* persistence in the lung [6], so deleting or modifying them to avoid infection may not be an optimal strategy in this environment. Given these factors, it is quite possible that *P. aeruginosa* CRISPR diversity exists within some cystic fibrosis patients. Our sample only contained four patients with detectable *P. aeruginosa* CRISPRs; a larger sample size may find lung environments with strains that have diversified their CRISPRs. Additionally, many other bacterial species which contain CRISPRs inhabit the lung; in just two patients, we identified 19 non-*Pseudomonas* CRISPR repeats from 14 different bacterial genera, including *Rothia* and *Streptococcus*.

These bacteria may each have their own types of interactions with phages, levels of reliance on CRISPR immunity versus other resistance mechanisms, and levels of CRISPR diversity in the population. Efforts to characterize the loci associated with these repeats and any diversity they might harbor are ongoing, and very preliminary results indicate that some samples do have species with multiple coexisting CRISPR alleles (Samantha DeWerff, personal communication). We are optimistic that among these we will find species in which we can observe the effects of distributed immunity in a natural microbial population.

Protospacer Typing: Leveraging Spacer Diversity to Track Phage Ecology

After observing the variation in *P. aeruginosa* CRISPR spacers between cystic fibrosis patients, we were inspired to assess diversity levels in the global population of *P. aeruginosa*. We found an incredible array of spacers, with highly diverse spacer content in strains from both cystic fibrosis patients and other environments around the world. This diversity reflects the wide variety of phage infecting *P. aeruginosa*, and we discovered that these spacers can be used to "protospacer type" phages, allowing us to identify and track which phage types these strains have encountered, providing ecological insight into phage distribution and migration. Thus far, we have only used protospacer typing on *P. aeruginosa*; however, this method should prove viable in any microbial system where the host has CRISPRs which actively sample phages or other elements of interest. In *P. aeruginosa*, protospacer typing confirmed panmictic host and phage populations; in systems which display stronger host biogeographic patterns or environmental differentiation, protospacer typing could be used to investigate if phages or mobile elements follow these patterns as well.

In addition to investigating phage ecology, there are numerous exciting applications for protospacer typing; for example, bacterial strains present in an infection could be screened for CRISPR spacers using spacer libraries matching possible phage therapy cocktails, allowing practitioners to choose the most effective treatment on a personalized level. This method could also be used to quickly identify and group prophages or other integrated mobile elements sampled by CRISPRs. Such elements are often of great interest because they can carry advantageous genes such as virulence factors, antibiotic

resistance genes, and genes encoding new metabolic functions, and protospacer typing provides the ability to rapidly identify and classify both well-studied and novel elements.

References

[1] Aagaard, K., Riehle, K., et al. (2012). A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One 7* (6), e36466.

[2] Held, N.L., Herrera, A., et al. (2010). CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS One 5* (9), e12988.

[3] Hraiech, S., Brégeon, F., et al. (2015). Bacteriophage-based therapy in cystic fibrosis-associated *Pseudomonas aeruginosa* infections: rationale and current status. *Drug Des. Devel. Ther. 9* , 3653–3663.

[4] James, C.E., Davies, E.V., et al. (2015). Lytic activity by temperate phages of *Pseudomonas aeruginosa* in long-term cystic fibrosis chronic lung infections. *ISME J. 9* (6), 1391–1398.

[5] Labrie, S.J., Samson, J.E., et al. (2010). Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol. 8* (5), 317–327.

[6] O'Toole, G.A. and Kolter, R. (1998). Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. *Mol. Microbiol. 30* (2), 295–304.

[7] Romero, R., Hassan, S.S., et al. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome 2* (1), 4.

[8] Walther-António, M.R.S., Jeraldo, P., et al. (2014). Pregnancy's Stronghold on the Vaginal Microbiome. *PLoS One 9* (6), e98514.

APPENDIX A

Tables

Table A.1. Complete List of *P. aeruginosa* Phage Clusters.

Phage Name	Phage Cluster	Protospacer Cluster
PAO1_627_663	01	4
PA14_643_675	01	4
PA7_4868_4912	01	15
PA7_671_781	01	15
LESB58_617_638	01	4
39016_575_650	01	4
M18_617_636	01	4
NCGM2.S1_5520_5596	01	4
DK2_620_639	01	4
B136-33_623_643	01	4
RP73_618_637	01	Unclustered
PA1_4213_4248	01	4
MTB-1_627_664	01	4
LES431_619_640	01	4
SCV20265_625_644	01	4
PA14_2245_2363	02	6
PA7_2282_2462	02	10
LESB58_2329_2530	02	6
39016_2947_3027	02	9
39016_3135_3235	02	6
39016_3305_3399	02	6
NCGM2.S1_3802_3895	02	6
DK2_2068_2214	02	6
B136-33_2187_2284	02	6
RP73_2844_3040	02	6
MTB-1_2189_2286	02	6
LES431_2276_2477	02	6
SCV20265_2319_2506	02	6
DMS3	03	2
MP22	03	2
MP29	03	2
MP38	03	2
LESB58_1604_1659	03	2
39016_5273_5326	03	2
PA1/KOR/2010	03	2
MP42	03	2

Table A.1. (cont.)

Phage Name	Phage Cluster	Protospacer Cluster
D3112	03	2
NCGM2.S1_5372_5424	03	2
PaMx73	03	2
LES431_1548_1603	03	2
JD024	03	2
PAO1_734_745	04	5
PA14_3998_4010	04	5
PA7_5142_5215	04	5
LESB58_4251_4263	04	5
M18_4418_4431	04	5
M18_4883_4901	04	5
B136-33_4385_4398	04	5
MTB-1_5097_5111	04	5
LES431_4198_4210	04	5
SCV20265_5267_5284	04	5
Pf1	04	5
PB1	05	3
LBL3	05	3
LMA2	05	3
14-1	05	3
SN	05	3
JG024	05	3
NH-4	05	3
PaMx13	05	3
KPP12_DNA	05	3
SPM-1	05	3
F8	05	3
PT5	06	7
LKD16	06	19
LUZ19	06	7
PT2	06	7
phikF77	06	19
phiKMV	06	7
vB_Pae-TbilisiM32	06	7
MPK7	06	7
MPK6	06	7
MBL	06	7
LESB58_2583_2651	07	1
39016_2143_2228	07	1

Table A.1. (cont.)

Phage Name	Phage Cluster	Protospacer Cluster
vB_PaeS_PMG1	07	1
PA1_1737_1784	07	1
LES431_2530_2598	07	1
SCV20265_1323_1386	07	1
D3	07	1
39016_5137_5191	08	2
NCGM2.S1_4860_4914	08	2
JBD18	08	2
JBD25	08	2
JBD67	08	2
B3	08	2
SCV20265_2698_2750	08	2
LESB58_1360_1425	09	1
LESB58_800_862	09	1
M18_1344_1415	09	1
PA1_1496_1566	09	1
MTB-1_4320_4398	09	1
LES431_1304_1369	09	1
F10	09	1
JG004	10	8
PaP1	10	8
PaMx12	10	8
vB_PaeM_C2-10_Ab1	10	8
PAK_P2	10	8
PAK_P4	10	8
PAK_P1	10	8
LUZ24	11	Unclustered
MR299-2	11	Unclustered
PaP3	11	99
vB_PaeP_C1-14_Or	11	Unclustered
vB_PaeP_p2-10_Or1	11	Unclustered
phiIBB-PAA2	11	Unclustered
TL	11	Unclustered
PA7_4273_4372	12	9
DK2_4747_4840	12	9
B136-33_4708_4807	12	9
RP73_983_1081	12	9
PA1_4626_4723	12	9
MTB-1_4864_4955	12	9
vB_Pae-Kakheti25	13	Unclustered

Table A.1. (cont.)

Phage Name	Phage Cluster	Protospacer Cluster
PaMx10	13	Unclustered
PaMx42	13	6
vB_PaeS_SCH_Ab26	13	1
73	13	21
CHA_P1	14	3
PAK_P5	14	3
P3_CHA	14	3
PAK_P3	14	3
KPP10	14	3
LIT1	15	12
PA26	15	12
vB_PaeP_C2-10_Ab09	15	12
PA7_141_160	16	Unclustered
PA7_45_116	16	4
39016_4_77	16	17
PA7_2684_2749	17	14
39016_2751_2807	17	14
PA1_3627_3647	17	Unclustered
phi297	18	1
NCGM2.S1_2862_2938	18	1
YMC/01/01/P52_PAE_BP	18	1
phiCTX	19	11
MTB-1_5610_5659	19	11
SCV20265_5782_5829	19	11
YuA	20	13
MP1412	20	13
M6	20	13
PaMx32	21	Unclustered
PaP2	21	Unclustered
119X	21	Unclustered
LESB58_4626_4666	22	23
LES431_4573_4600	22	Unclustered
DK2_2836_2922	23	11
SCV20265_3070_3116	23	5
39016_819_882	24	18
NCGM2.S1_1519_1583	24	18
H66	25	10
F116	25	10
PAJU2	26	1

Table A.1. (cont.)

Phage Name	Phage Cluster	Protospacer Cluster
MTB-1_3073_3155	26	1
39016_6005_6024	27	20
NCGM2.S1_5852_5868	27	20
PA14_1790_1819	28	Unclustered
39016_3829_3856	28	Unclustered
M18_2482_2554	29	10
RP73_2687_2759	29	10
PaMx28	30	16
PaMx74	30	16
phiKZ	31	Unclustered
PA7	31	Unclustered
phiEL	32	1
PA7_350_367	33	5
PP7	34	Unclustered
PA1_5723_5781	35	11
Pf3	36	Unclustered
B136-33_1185_1197	37	22
PaMx25	38	12
PaMx11	39	1
PaMx31	40	3
PaBG	41	2
PhiPA3	42	Unclustered
PA7_4959_4986	43	17
PA11	44	3
LUZ7	45	4
LKA1	46	Unclustered

Table A.2. Protospacers unique to a single phage genome cluster (cluster-defining, intermediate, or individual). Clusters without protospacers not shown.

Cluster	Cluster-Defining	Intermediate	Individual
cluster01		24, 263, 319, 496, 576, 658, 683, 695, 746, 843, 855, 956, 972, 1070, 1130, 1131, 1158, 1228, 1311, 1312, 1417, 1428, 1631, 1703, 2083, 2305, 2359, 2360, 2392, 2402, 2545, 2551, 2657, 2740, 2761, 2775, 2826, 2852, 3024	1245, 1251, 1554, 1831, 2258, 2440, 2790, 3071, 3080
cluster02		69, 237, 254, 408, 486, 516, 1757, 1941, 2765, 2824, 2835, 3014	52, 65, 72, 129, 149, 175, 218, 241, 336, 365, 440, 450, 480, 531, 538, 557, 599, 654, 655, 754, 807, 844, 948, 949, 974, 978, 982, 1003, 1095, 1161, 1168, 1203, 1238, 1244, 1340, 1432, 1697, 1806, 1812, 1824, 1895, 1929, 2070, 2228, 2231, 2236, 2244, 2273, 2299, 2374, 2385, 2485, 2533, 2568, 2583, 2825, 2844, 2860, 2885, 2886, 2889, 2941, 2944

Table A.2. (cont.)

Cluster	Cluster-Defining	Intermediate	Individual
cluster03	180, 182, 209, 216, 228, 358, 453, 1189, 1590, 1596, 1789, 2195, 2240, 2498, 2541, 2586, 2694, 2704, 2780, 2792	44, 56, 102, 181, 184, 185, 210, 233, 242, 243, 245, 259, 270, 271, 302, 303, 344, 345, 407, 426, 488, 489, 503, 504, 584, 585, 586, 587, 603, 608, 633, 706, 769, 781, 814, 834, 842, 858, 950, 984, 1047, 1050, 1057, 1108, 1111, 1119, 1184, 1194, 1280, 1295, 1400, 1420, 1453, 1457, 1465, 1471, 1478, 1491, 1564, 1572, 1595, 1670, 1692, 1693, 1786, 1813, 1879, 1897, 1898, 1917, 1919, 1920, 1924, 1931, 1934, 1935, 1936, 2000, 2112, 2174, 2267, 2272, 2290, 2435, 2486, 2668, 2751, 2753, 2781, 2791, 2803, 2880, 2884, 2914, 2915, 2918, 2932, 2948, 2951, 3018, 3078, 3079, 3083, 3105, 3140	1659, 1930, 1990, 2949
cluster04	130, 435, 465, 876, 1858, 1903, 1906, 1908, 2057	127, 406, 1850, 1859, 1905, 1909, 2723, 2724, 2728, 2729, 2731, 2732, 2734, 2735, 2736, 2744, 2745, 2748, 2749, 3049, 3050, 3051, 3055, 3056, 3062, 3063	31, 146, 634, 1167, 1907, 2746
cluster05	2957	84, 1192, 2821	
cluster06		423, 1365, 2366	

Table A.2. (cont.)

Cluster	Cluster-Defining	Intermediate	Individual
cluster07		59, 143, 208, 308, 540, 549, 622, 628, 671, 798, 818, 831, 847, 958, 959, 1069, 1230, 1248, 1315, 1385, 1505, 1508, 1588, 1632, 1665, 1748, 1775, 1815, 1978, 2079, 2089, 2136, 2153, 2154, 2183, 2199, 2229, 2304, 2319, 2320, 2472, 2684, 2805, 2849, 2862, 2865, 2866, 2867, 2868, 2910, 2936	21, 160, 193, 306, 415, 659, 677, 690, 762, 780, 782, 792, 945, 1114, 1115, 1225, 1319, 1352, 1456, 1507, 1597, 1634, 1819, 1828, 1951, 2381, 2382, 2426, 2430, 2462, 2595, 2764, 2773, 2830, 2840, 2938, 2964, 2984, 3130
cluster08	1738, 2082	98, 162, 172, 173, 249, 348, 359, 367, 368, 380, 392, 422, 446, 482, 533, 597, 609, 681, 704, 767, 799, 809, 832, 888, 922, 934, 968, 1055, 1138, 1163, 1199, 1212, 1279, 1287, 1474, 1562, 1714, 1715, 1739, 1773, 1838, 1845, 1851, 1852, 1853, 1961, 1976, 1977, 1995, 1996, 1997, 1998, 2023, 2269, 2300, 2332, 2396, 2453, 2481, 2518, 2585, 2613, 2614, 2623, 2629, 2795, 2804, 2854, 2906, 3046, 3111, 3116, 3118, 3132	315, 366, 541, 630, 990

Table A.2. (cont.)

Cluster	Cluster-Defining	Intermediate	Individual
cluster09	64, 556, 562, 666, 714, 761, 778, 981, 983, 1010, 1246, 1448, 1684, 2051, 2519, 2738	46, 78, 91, 103, 139, 157, 187, 320, 356, 357, 434, 454, 471, 591, 619, 645, 652, 679, 751, 764, 841, 879, 928, 941, 944, 971, 997, 1032, 1033, 1037, 1133, 1190, 1211, 1218, 1416, 1472, 1570, 1737, 1756, 1835, 1894, 2095, 2115, 2175, 2219, 2220, 2330, 2352, 2474, 2483, 2508, 2512, 2538, 2539, 2540, 2553, 2571, 2647, 2848, 2878, 3028, 3138	55, 457, 487, 546, 570, 571, 775, 862, 1100, 1128, 1350, 1512, 1620, 1746, 1847, 2041, 2061, 2092, 2137, 2139, 2208, 2227, 2256, 2383, 2399, 2493, 2552, 2669, 2670, 2774, 2853, 2859, 2966
cluster10		427	
cluster12	8, 192, 400, 811, 835, 859, 1117, 1118, 1288, 1628, 1702, 1882, 2158, 2355, 2361, 2511, 2768, 2769, 2892, 2943	19, 647, 860, 1012, 1013, 1281, 1630, 1639, 1705, 1950, 2243, 2266, 2556, 2632, 3077	2685, 2959
cluster13			2386, 2935
cluster14	502	863	
cluster15			3
cluster16			1068
cluster17			827, 873, 2363, 2873, 2940
cluster18	287, 322, 347, 402, 416, 513, 728, 897, 899, 1175, 1645, 1777, 1964, 1981, 2030, 2049, 2050, 2052, 2424, 2438, 2497, 2560, 2603, 2604, 2608, 2996	251, 355, 360, 468, 921, 1198, 1296, 1347, 1656, 1699, 1704, 2012, 2087, 2415, 2484, 2606, 2618, 2783, 3149	7, 235, 252, 383, 515, 870, 901, 986, 1110, 1351, 1397, 1475, 1494, 1728, 1889, 1940, 1980, 2053, 2121, 2255, 2406, 2605, 2607, 2963, 2992, 3020

Table A.2. (cont.)

Cluster	Cluster-Defining	Intermediate	Individual
cluster19	152, 154, 199, 264, 346, 393, 394, 451, 572, 574, 579, 580, 614, 692, 732, 889, 914, 926, 927, 992, 993, 999, 1039, 1080, 1107, 1164, 1176, 1221, 1727, 1809, 1866, 1878, 1933, 2062, 2162, 2176, 2197, 2234, 2329, 2347, 2590, 2616, 2619, 2630, 2634, 2778, 2847, 2969	106, 107, 220, 333, 560, 910, 915, 975, 1092, 1575, 1910, 1911, 1948, 2177, 2505	425, 568, 848, 1278, 1576, 2097, 2620, 2621, 2971
cluster20			1234
cluster22			2558
cluster23			4, 39, 136, 137, 151, 176, 191, 194, 195, 200, 206, 291, 317, 327, 405, 410, 412, 462, 490, 500, 532, 578, 702, 707, 725, 755, 839, 849, 874, 890, 898, 907, 979, 1000, 1104, 1105, 1122, 1135, 1148, 1155, 1157, 1214, 1215, 1233, 1242, 1266, 1335, 1348, 1359, 1373, 1408, 1409, 1433, 1438, 1477, 1479, 1520, 1521, 1541, 1549, 1565, 1587, 1608, 1609, 1616, 1687, 1698, 1707, 1747, 1849, 1857, 2044, 2069, 2072, 2182, 2185, 2194, 2209, 2274, 2292, 2293, 2316, 2495, 2600, 2635, 2643, 2687, 2688, 2716, 2720, 2742, 2800, 2841, 2842, 2843, 2922, 2928, 2968, 2979, 2989, 3006, 3008, 3009, 3010, 3011, 3032, 3033, 3059, 3072, 3102, 3141, 3151
cluster24	253, 595, 668, 1872		

Table A.2. (cont.)

Cluster	Cluster-Defining	Intermediate	Individual
cluster25	279, 280, 511, 512, 1106, 1556, 1721, 1793, 2337, 2353, 2857		123, 285, 293, 476, 711, 712, 713, 745, 846, 1082, 1083, 1084, 2004, 2534, 2535, 2939
cluster26	1, 70, 81, 85, 469, 891, 906, 1014, 1284, 1379, 1452, 1594, 1700, 1861, 1862, 2016, 2171, 2172, 2201, 2490, 2548, 2712, 2798, 2829		17, 40, 41, 42, 87, 120, 150, 203, 204, 229, 328, 332, 337, 375, 387, 430, 452, 473, 491, 565, 593, 626, 669, 675, 708, 736, 757, 758, 759, 765, 770, 923, 947, 988, 1021, 1034, 1062, 1088, 1123, 1134, 1196, 1209, 1217, 1222, 1243, 1285, 1290, 1354, 1380, 1410, 1422, 1439, 1467, 1527, 1530, 1561, 1591, 1637, 1647, 1655, 1711, 1712, 1719, 1730, 1732, 1767, 1779, 1788, 1801, 1805, 1860, 1921, 1963, 2025, 2038, 2055, 2066, 2099, 2102, 2108, 2122, 2123, 2132, 2133, 2135, 2155, 2160, 2181, 2200, 2205, 2211, 2214, 2216, 2241, 2246, 2254, 2259, 2276, 2314, 2351, 2370, 2384, 2431, 2434, 2461, 2480, 2504, 2509, 2513, 2514, 2529, 2530, 2594, 2664, 2713, 2785, 2796, 2801, 2827, 2869, 2877, 2985, 3082

Table A.2. (cont.)

Cluster	Cluster-Defining	Intermediate	Individual
cluster29	105, 144, 156, 188, 442, 771, 911, 1539, 1569, 1577, 1578, 1579, 1914, 1915, 1923, 1942, 1956, 2163, 2233, 2311, 2328, 2334, 2389, 2391, 2395, 2409, 2418, 2420, 2421, 2422, 2423, 2442, 2443, 2444, 2445, 2446, 2448, 2450, 2506, 2507, 2699, 2831, 2832, 2833, 3027, 3084, 3115, 3134		942, 2029
cluster30	1149		742, 1975, 2482